Faculty of Engineering and Architecture
Departement ELIS
2015–2016

Recognize Emotion in the brain using EEG Data

by

Andreas DE LILLE

Promotors: Prof. J. DAMBRE
Dr. Ir. P. VAN MIERLO
Assistent: Ir. T. VERHOEVEN

Contents

Nomenclature

BCI	Brain Computer Interface
CSP	Common Spatial Patterns
DASM	Differential Asymmetry
DCAU	Differential Caudality
DEAP	Dataset for Emotion Analysis using Physiological Signals
EEG	Electroencephalography
GSR	Galvanic Skin Response
LDA	Linear Discriminant Analysis
MEG	magnetoencephalography
OCR	Optical Character Recognition
PSD	Power Spectral Density
RASM	Rational Asymmetry
RCAU	Rational Caudality
SAM	self-assessment manikins

.

1

Introduction

This chapter describes the context of the thesis, starting with brain computer interfaces(BCI), before defining some BCI basics. After that, the P300 speller and P300 paradigm are introduced. Before the need for an emotionally aware P300 speller is justified, the basic process of emotion in the brain is explained.

1.1 Brain computer interfaces

A Brain Computer Interface (BCI), creates a direct neural link from the brain to the computer[1], that tries to recognize patterns and based on the extracted information, performs actions. A BCI removes to need for physical actions, i.e. typing or moving a mouse, for the transfer of information. The neural link provided by the BCI is made of two important components. The first component is the extraction component, which extract brain signals from the brain. The second component is the computer that interprets signals and performs actions based on the outcome.

1.1.1 Electroencephalography (EEG)

Different technologies exist to analyze the brain, the most convenient method is via Electroencephalography (EEG), since it is a non-invasive method. Non-invasive methods, in contrast to invasive methods require no surgery; they simply measure electrical activity using electrodes placed on the scalp.

The electrical activity in a brain is caused when an incoming signal arrives in a neuron. This triggers some sodium ions to move inside the cell, which in turn, causes a voltage rise[2]. When this increase in voltage reaches a threshold, an action potential is triggered in the form of a wave of electrical discharge that travels to neighboring neurons. When this reaction occurs simultaneously in a lot of neurons, the change in electrical potential becomes significantly, making it visible to the EEG surface electrodes. EEG can thus only capture synchronized activity of many, many neurons.

Signals originating from the cortex, close to the skull, are most visible, while signals originating deeper in the brain cannot be observed directly. Even for signals originating close to the cortex, EEG is far from precise as the bone between the the cortex and electrodes distorts the signal.

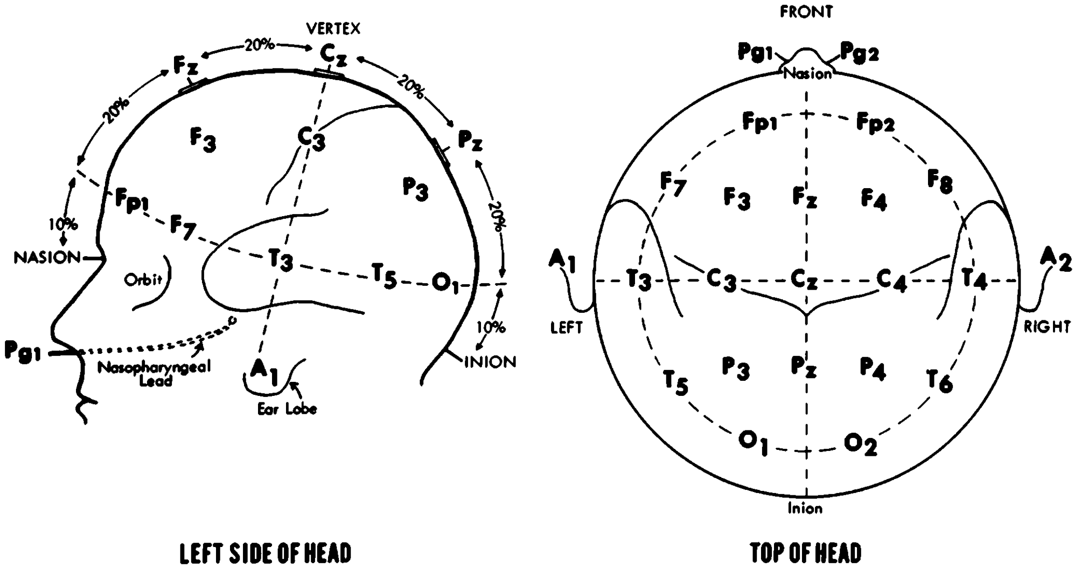
Additionally other artifacts like eye and muscle movement add a lot a noise to the signal, noise removal techniques are therefor advised. Even though the noise is persistent and EEG data has very low spatial resolution, it still can provide significant insight into the electrical activity of the cortex while offering excellent temporal resolution[3].

Note that EEG records electrical activity, other methods like magnetoencephalography (MEG) measure brain activity using magnetic fields. Since MEG is more prone to noise from external magnetic signals, i.e. the earth's magnetic field and electromagnetic communication, a magnetic shielded room is required, making this method very expensive and not mobile.

EEG uses electrodes which are placed on the scalp to measure the electrical activity. To ensure that experiments are replicable, standards for locations of electrodes have been developed. One of these systems is the 10/20 system, an internationally recognized methods to describe location of scalp electrodes[4]. The numbers 10 and 20 refer to the distances between the electrodes, which are either 10% or 20% of the total front-back or left-right distance of the skull. Each site is identified with a letter that determines the lobe and hemisphere location.

- **F:** Frontal
- **T:** Temporal
- **C:** Central
- **P:** Parietal
- **O:** Occipital

Note that no central lobe exists; the C letter is only used for identification purposes. The letter z indicates that the electrode is placed on the central line. Even numbers are use for the right hemisphere, while odd numbers are used for the left hemisphere. A picture of a 23 channel 10/20 system is added below for clarification. Even though some experiment setups may use a different set of channels than shown in figure 1.1, they all follow the same naming convention.



LEFT SIDE OF HEAD

TOP OF HEAD

Figure 1.1: The electrode placement of a 23 channel system[5].

Two different types of EEG channels exist, monopolar and dipolar. A monopolar channel records the potential difference of a signal, compared to a neutral electrode, usually connected to an ear lobe of mastoid. A bipolar channel is obtained by subtracting two monopolar EEG signals, which improves SNR by removing shared artifacts[6].

In the frequency domain, brain waves are usually split up into different bands[7, 8], each band has a different medical interpretation. These wavebands are:

1. **Alpha:** 8-13Hz, indicate how relaxed and/or inactive the brain is.
2. **Beta:** 13-30HZ, indicate a more active and focused state of mind.
3. **Gamma:** 30-50Hz, relate to simultaneous processing of information from different brain areas.
4. **Delta:** 0-4hz, these waves are generated during dreamless sleep and meditation.
5. **theta:** 4-8Hz, occur during dreaming.

Most muscle and eye artifacts have a frequency around 1.2Hz. Artificats caused by nearby power lines, have a frequency around 50Hz[2]. To remove most of this noise, a bandpass filter is usually applied to filter out frequencies below 4Hz and above 40-45Hz.

1.2 Emotion recognition

Psychology makes a clear distinction between physiological behavior and the conscious experience of an emotion, called expression[2]. The expression consists of many parts, including the facial expression, body language and voice concern. Unlike expression, the physiological aspect of an emotion, e.g. heart rate, skin conductance and pupil dilation, is much harder to control. To really know one's emotions, it seems, one has to research the physiological aspect of the emotion. One possibility for this is analysis of brain activity via Electroencephalography[9], which is the main method for this thesis.

1.2.1 Emotion in the brain

Before emotions can be recognized, a classification model is needed. A common model to classify emotions is the bipolar arousal-valence model[2, 10], that places emotions in a two dimensional space. The main advantage of using a multidimensional model, is that all emotions are modelled in its space, even when no particular discrete label can be used to define the current feeling. Figure 1.2 shows the mapping of different emotions for this model.

Even though arousal and valence describe emotion quite well, a third dimension can also be added. The new model then has three dimensions: arousal, valence and dominance. Arousal indicates how active a person is and ranges from inactive, bored to active, excited. The valence indicates if the emotion is perceived as positive or negative. The third dimension, the dominance, indicates how strong the emotional feeling was and ranges from a weak feeling to an empowered, overwhelming feeling. The dominance component can aid to filter out samples of strong feelings, since feelings with low dominance are less likely to show significant effects.

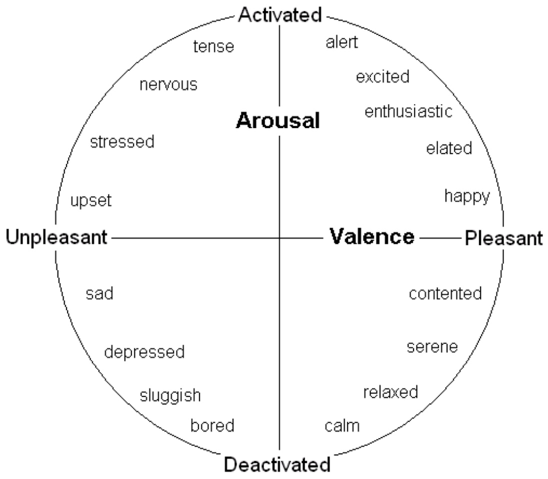


Figure 1.2: The arousal - valence model maps emotions in a two dimensional plane.

1.2.2 Features

The next step is defining relevant physiological features for emotion recognition, two categories of features are observed: EEG features and non-EEG features. The EEG features are features extracted from the electroencephalography measurements from the subject's scalp. This section will go through the used EEG features in this thesis.

The power spectral density (PSD) of a signal gives the distribution of the signal's energy in the frequency domain. By calculating the spectral density for different waveband of the signal, one can determine how much alpha, beta, ... power is in the signal.

Differential entropy is defined as follows [11]

$$-\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp(\frac{(x-\mu)^2}{2\sigma^2}) log(\frac{1}{\sqrt{2\pi\sigma^2}}) exp(\frac{(x-\mu)^2}{2\sigma^2}) dx$$

According to [12], the differential entropy of a certain band is equivalent to the logarithmic power spectral density for a fixed length EEG sequence, which simplifies the calculations significantly.

The most known feature for valence recognition is the frontal asymmetry of the alpha power[3]. The right hemisphere is generally speaking, more active during negative emotion than the left hemisphere which is in turn more active during positive emotions[10, 9, 11]. The asymmetry can be calculated in different ways, one of them is the differential asymmetry (DASM) , where the left alpha power is subtracted from the right alpha power.

$$DASM = DE_{left} - DE_{right}$$

Another way to measure the asymmetry if by division. The Rational Asymmetry (RASM) does exactly this and is given by:

$$RASM = \frac{DE_{left}}{DE_{right}}$$

With DE_{left} and DE_{right} being the left and right differential entropy respectively.

Another reported feature in literature is the caudality, or the asymmetry in fronto-posterior direction[13]. This can again be calculated in two ways. The first method is the differential Caudality (DCAU) is defined as:

$$DCAU = DE_{front} - DE_{post}$$

Another way to determine the Caudality is the Rational Caudality (RCAU) , which is defined as:

$$RCAU = \frac{DE_{front}}{DE_{post}}$$

With DE_{front} and DE_{post} being the frontal and posterior power respectively.

One way to determine the arousal is by looking at the different wavebands. Each waveband has their own medical interpretation, see 1.1.1. More alpha power corresponds to a more relaxed brain, while more beta power corresponds to a more active brain. The alpha / beta ratio therefore seems a good indicator for the arousal state of a person.

The aforementioned EEG features are just one class of physiological features. The DEAP dataset contains several physiological measurements, listed below [14]. For each of these measurements the average and the standard deviation is calculated.

The Galvanic Skin Response uses two electrodes on the middle and index finger of the subjects left hand to measure the skin resistance. It has been reported that the mean value of the GSR is related to the level of arousal[15, 14].

The respiration belt, indices the user's respiration rate. Slow respiration is linked to relaxation (low arousal), while fast and irregular respiration patterns corresponds to anger or fear, both emotions with low valence and high arousal[14].

A plethysmograph is a measurement of the volume of blood in the subject's left thumb. This can be interpreted as the the blood pressure. Blood pressure offers valuable insight into the emotional state of a person as it correlated with emotion; stress is known to increase blood pressure[14].

The heart rate is not explicitly in the DEAP dataset, but can be extracted from the plethysmograph, by looking at local minima and maxima[14]. This is clearly visible when looking at the plethysmograph's output, shown in Figure 1.3.

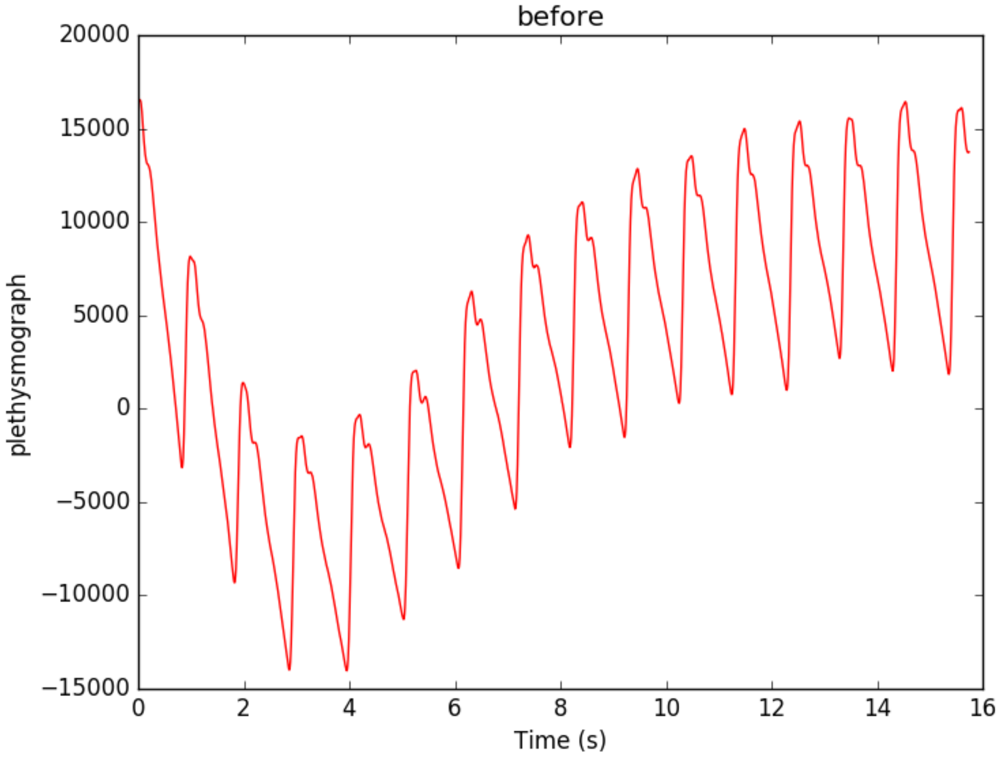


Figure 1.3: The plethysmograph before smoothing.

The heart rate extraction is done in two steps. First the plethysmograph signal is smoothed using a butter filter to avoid noise being selected as local optima. In the second stage the local optima are located, which is shown in Figure 1.4

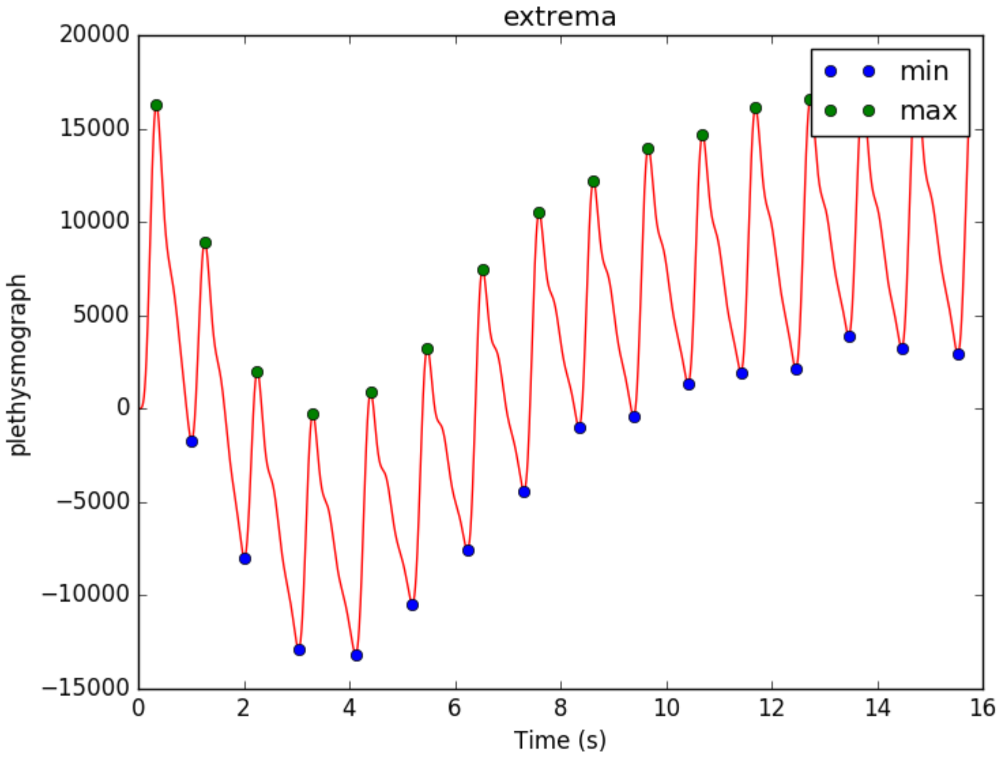


Figure 1.4: The local optima in the plethysmograph.

These optima correspond to a heart beat, therefore the time between two consecutive local minima or maxima corresponds to the time between two heart beats, known as the interbeat interval. Getting the average heart rate from the interbeat interval is straight forward.

The last physiological feature is the skin temperature of the subject.

1.3 Machine learning

The next important topic that needs to be covered is machine learning. Machine learning is the missing link between the features and the emotion recognition. As machine learning is a very broad domain, the discussion will be limited to the application of machine learning and machine learning techniques as this is the most relevant part for this thesis. One possible definition for Machine learning is: "the science of getting computers to act without being explicitly programmed". To do so, machine learning uses pattern recognition to find patterns or structure in the data. A simple example of machine learning is the Optical Character Recognition (OCR), where a computer recognises characters in pictures.

Let's get a look at the following example, to further explain how machine learning works. Suppose one has a price list of houses that are for sale combined with their total area. Logic sense dictates us that a bigger house will have a higher asking price than a smaller house. Therefore the asking price of a house is correlated to the asking price. Suppose you want to predict how much a certain home is worth, based on their area. This is possible with machine learning, first you need to train your machine learning algorithm with a list of asking prices and the corresponding area of the house. This should give you a coefficient, lets say you pay 1000 euro for each square meter. Once this is done you can predict prices of new houses based on the corresponding area.

This will give some reasonable results, but the algorithm will probably have some flaws. This is due to the fact that the area of the house is only one feature that determines the price, there are many other that we haven't taken into consideration. Looking with more detail at the data, i.e. adding additional features will thus improve the performance of our algorithm. For example, a house with 5 bedrooms is more expensive than a house with only 3 bedrooms.

Machine learning algorithms are responsible for finding the relation between features and the predicted value. There exist many machine learning algorithms, one way to group these algorithms, is to look at their produced output. In the asking price examples above, the output is a price, which is a continuous value. The OCR example from above, where characters are recognized in a picture is a classification problem, as there is only a limited set of characters.

Another way to group the algorithms is based on their training data. In the asking price examples above one gets labelled results; the asking price is given for each area, this is referred to as supervised machine learning. The other possibility is unsupervised machine learning, which often results in finding groups of similar data points (clustering), without knowing the actual labels. Note that the combination of supervised and unsupervised data, also known as semi supervised learning, is also possible. Suppose you have a dataset with 5000 webpages and you want to categorise them in 10 distinct categories, e.g. science, nature, cooking, ... , but you only have the labels for 100 of the 5000 pages. Then you could first cluster the pages in similar groups using unsupervised learning. As soon as a group contains a labelled page, you can label all the pages in the group, since clustering returns groups of similar examples. Semi supervised learning has the advantage that one can also use unlabelled data. Unlabelled data is often easy and cheaper to obtain, unlike labelled data which is usually quite rare; if you had a fast and easy way to label the data then you wouldn't be needing machine learning.

In this thesis machine learning is used to find patterns in the aforementioned features that indicate the user's emotional state.

1.3.1 Over and underfitting / high bias and high variance

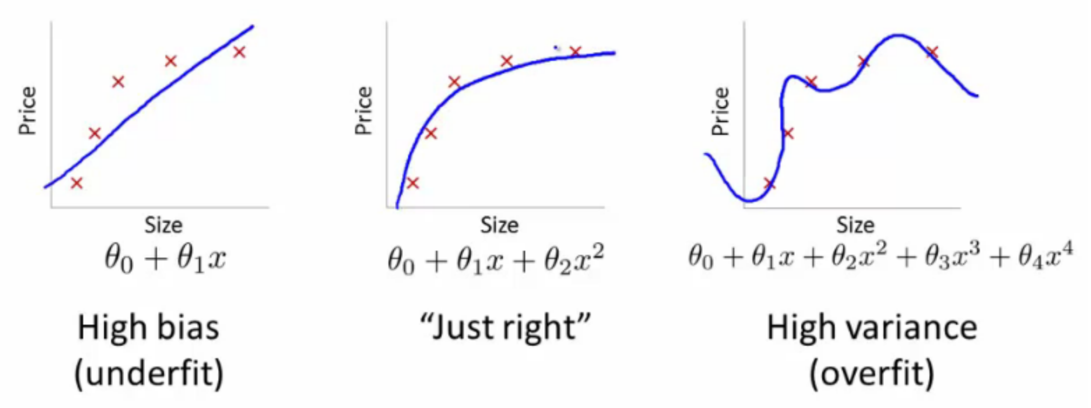


Figure 1.5: Overfitting versus underfitting[16].

Suppose the example in Figure 1.5, where one tries to find a good function to fit the given data points. Looking at the three proposed functions, one can easily see that the middle figure is the most likely generator function of the red points.

The figure on the left corresponds to an underfit, where the proposed function is not able to capture sufficient detail of the points. The function is not complex enough to approach the generator function, which is known as high bias. A high bias problem has a high training error, as the function is not able to fit the points correctly, this is visible in Figure 1.6

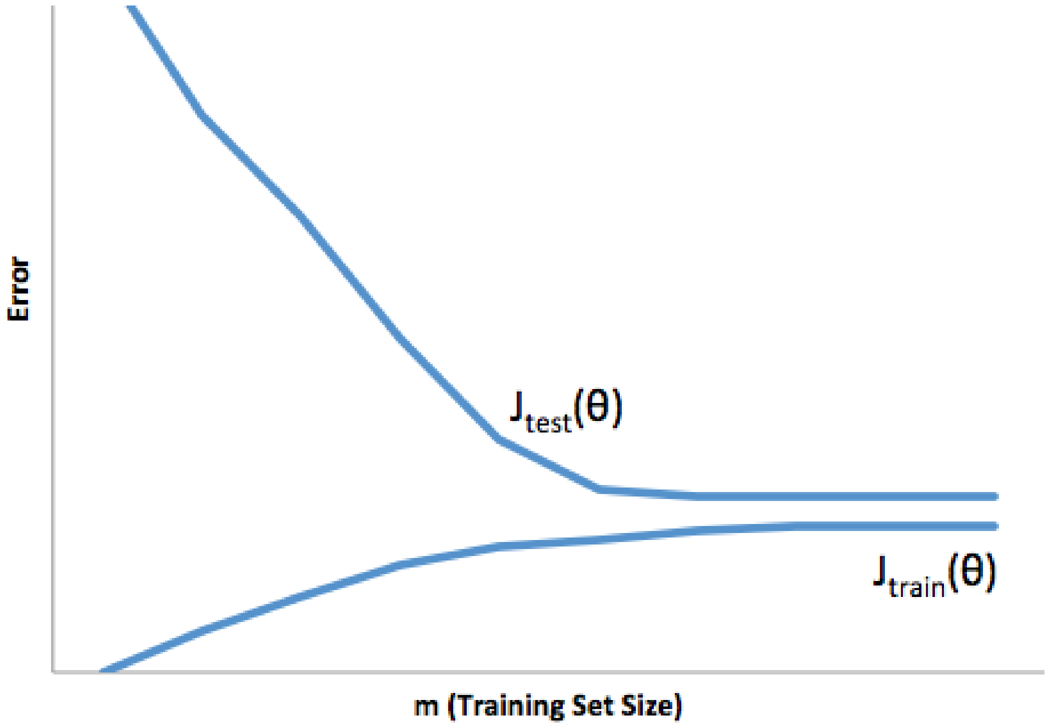


Figure 1.6: A high bias function is not complex enough to approach the generator function closely.

The function on the rights The function on the right corresponds to an overfit; the function 'goes through' each point exactly, but one can see that in between data points the behaviour of the hypothesis function is not logical. This problem is known as a high variance problem, where the train error is close to zero, but the test error is quite dramatic.

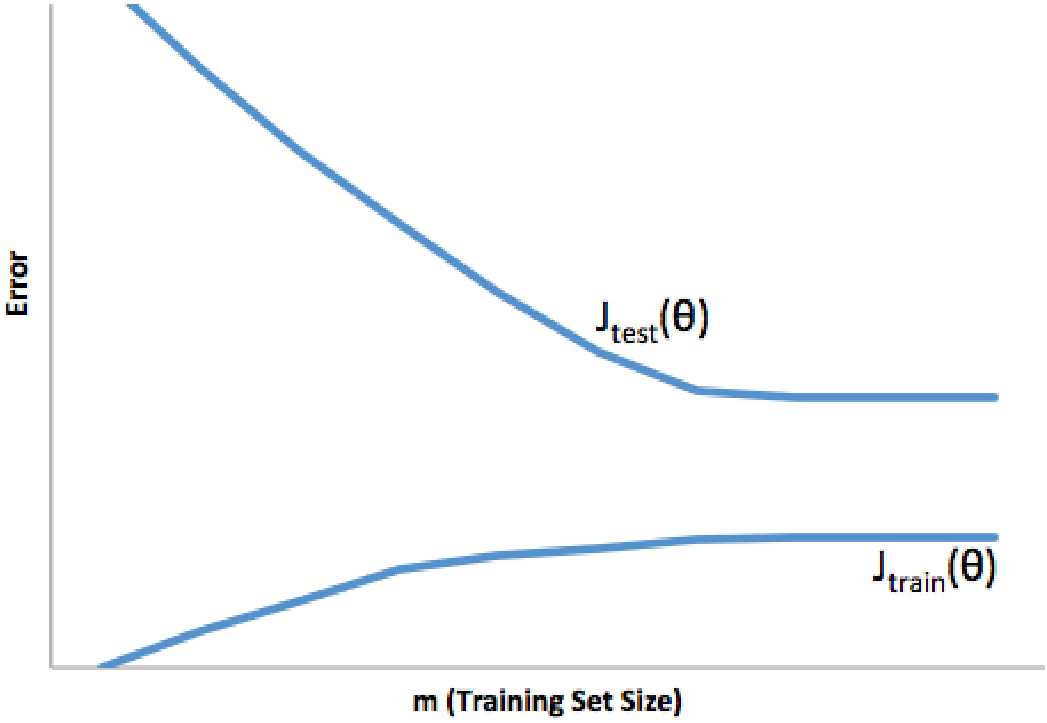


Figure 1.7: A High variance function is too complex and fits the data point too closely.

Another way to explain the bias variance tradeoff is by an example. Suppose you have a dart board, as shown in Figure 1.8. Suppose the situation on the top left corner, this corresponds to a world class player that has perfect aim, and very little variation on his precision. The situation on the left bottom corresponds to a player that has very little variation on his precision, but that is consistently aiming too high, he is biased to hit higher than needed. The pictures on the right side are different, there the person may or may not have a biased aim, but it is clear that he has a lot of variation in the precision of his aim.

In the context of machine learning, the low bias corresponds to having a hypothesis set that is close to the generator function, which allows you to get quite close. However you still have to pick the right function from that set, which is hard to do if you don't have enough data. If you are not able to take the best solution from the hypothesis set, you have high variance.

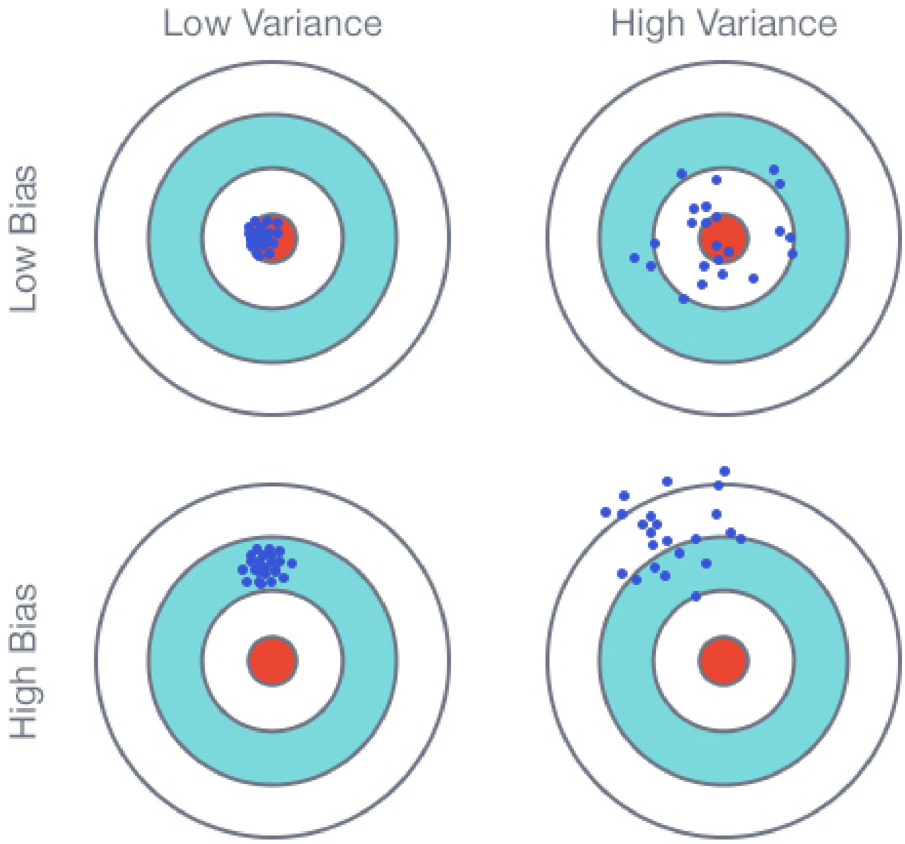


Figure 1.8: The bias variance explained using the dartboard example found at [17]

1.3.2 Support Vector Machines (SVM)

1.3.3 Random Forests (RF)

1.4 Goal of the thesis

The first goal is finding relevant features for emotion recognition in a person specific setting. This algorithm tries to accurately predict the emotions of one person, using only training data from that person. The Second goal is finding features for emotion recognition in a cross-persons setting. In this setting the features should generalise well across different persons, thus the algorithm should be able to recognize emotions from unseen persons.

The main problem is that there are already a lot of features known, but, as is often the case with EEG data, training data is expensive and limited. Using a lot of features will thus quickly result in overfitting.

Additionally, a lot of different features are reported in the literature, as you can see in Table **??**.

Table 1.1: Six different papers on emotion recognition, six different feature sets

study	**features used**
[?]	Alpha and beta power
[?]	PSD and asymmetry features
[?]	PSD
[?]	discrete wavelet transform of alpha, beta and gamma band
â [2]	alpha/beta ratio, Fpz beta and alpha band power
[11]	PSD, RCAU, DCAU, DASM, RASM, DE

Another point to note is that simply using a limited set of features, might work, but is likely to lead too less accurate results as optimal features might have been left out. This problem is even more severe when considering that EEG data is person specific, features that work good for one person, might not work for another person. Finding a good set of features that works for all persons is a non trivial problem. One solution to this problem could be to use a large pool of possible features from which a limited set of good features are selected. This allows to provide good features to the machine learning algorithm, while still keeping the set of features limited in size. The machine learning algorithm which features to use and which to neglect.

1.4.1 Datasets

One of the most used datasets in the context of emotion recognition is the Dataset for Emotion Analysis using Physiological Signals (DEAP) dataset[14]. This dataset contains EEG samples at 512 Hz of 32 persons each viewing 40 videos. A preprocessed version of this dataset, that is down sampled to 128Hz and has EOG removal will be used extensively during this thesis.

2

A first look at the data

2.1 The DEAP dataset

This thesis uses the DEAP dataset[14], a dataset for emotion analysis that is publicly available for academic research. This dataset contains EEG recordings of 32 participants, each watching 40 one minute excerpts of music videos. Each video was rated individually by each person on valence, arousal, dominance and liking. The first three ratings correspond to the valence, arousal and dominance space of an emotion 1.2.1. The liking component indicates how much the person liked the video excerpt and should not be confused with the valence component; it inquires information about the participants' tastes, not their feelings, i.e. a person can like a video that triggers angry or sad emotions. The liking rates are neglected, since they are not part of the emotion space.

For assessment of these scales, the self-assessment manikins (SAM), were used[14]. SAM visualizes the valence, arousal and dominance scale with pictures, each picture corresponds to a discrete value. The user can click anywhere in between the different figures, which makes the scales continuous. All dimension are given by a float between 1 and 9, but for the context of this thesis, a preprocessing step scaled and translated these values to ensure they range between 0 and 1.

The used SAM figures are shown in Figure 2.1. The first row gives the valence scale, ranging from sad to happy. The second row shows the arousal scale, ranging from bored to excited. The last row represents the different dominance levels. The left figure represents a submissive emotion, while the right figure corresponds with a dominant feeling.

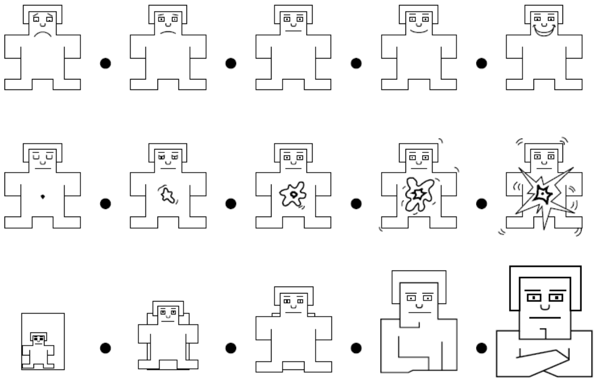


Figure 2.1: The images used for the SAM[14].

To further inspect the distribution of the user ratings and whether or not the data is balanced, the average for each emotion dimension (valence, arousal and dominance), was determined using all videos of all persons. These can be seen in Table 2.1. A uniform distribution, which is the ideal case for machine learning, should give a value of 0.5. The averages of the different dimensions are a just a little above 0.5, which gives a first indication that overall, the data is only slightly unbalanced.

	Valence	**Arousal**	**Dominance**
value	0.532	0.520	0.548

Table 2.1: The average value of each component.

2.2 A first model to classify the valence

The first model in this thesis recognizes the valence of a single person. This was done with a linear SVM classifier in a first attempt, since SVMs were used in a lot of the covered literature. SVMs have the advantage that they are able to handle limited datasets well, because they only look at points close to the separation boundary. In this attempt 2 features were compared, the usage of the frontal alpha asymmetry and the frontal theta power. Both feature are in literature frequently reported as being correlated to the valence. A third feature set used a combination of both these features.

The first classifier labels the valence values into two classes: low valence (unpleasant feelings) and high valence (pleasant feelings). This also meant that labels needed to be assigned to the dataset. This can be done in different ways.

The first and most straightforward method simply splits the valence range in half, all values positioned in the first and second half of the valence region were assigned to the low and high valences classes respectively. Note that even though the data set is overall quite balanced, the situation becomes quite different in case of a person specific classifier. Looking at Figure 2.2, it is clear that for some persons the data is quite unbalanced, for examples the unbalance is quite high for person 6. There are only 10 examples assigned to the low valance class, while 30 examples are assigned to the high valence class. Having an unbalance in training examples leads to less accurate classifiers.

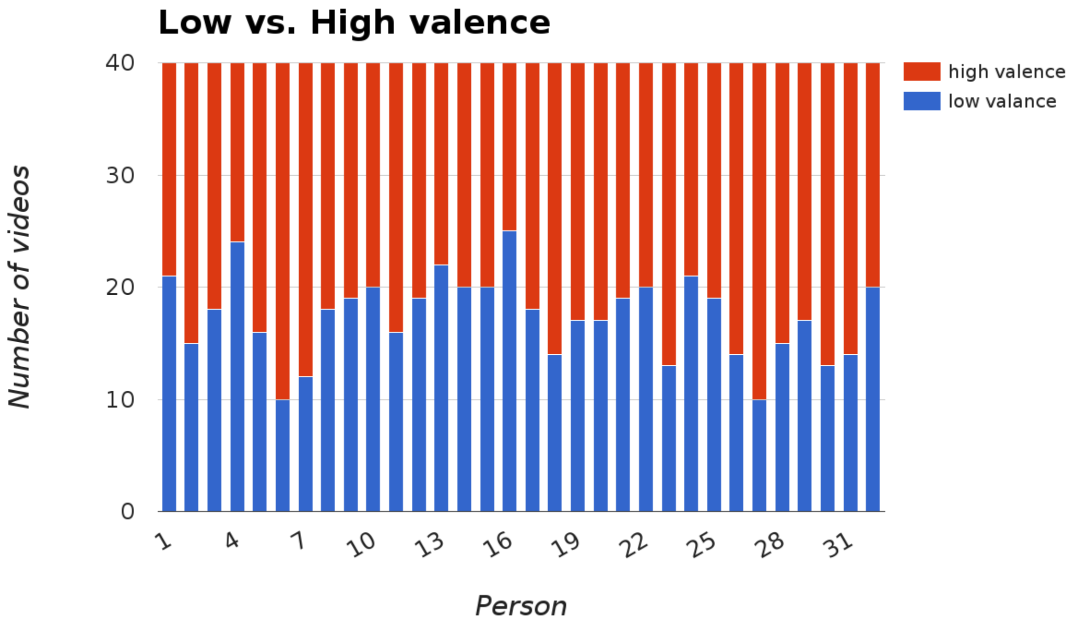


Figure 2.2: This graph show the assignment of the data set in to two classes: high and low valence. Each bar on the X axis represents a person, while the Y axis represents the number of low and high valence values.

The unbalance and especially the difference between unbalances for different persons is actually quite remarkable, given that each persons watched the same set of videos. Even taking into account personal differences, the different is high. One explanation for this might be that each persons rated movies differently; some persons are prone to giving higher values than other which results in an higher average valence value for some persons. To solve this problem one could simply order all the rating in ascending or descending order and assign the lower and higher half to the lower and higher valence class respectively. Using the median of the ratings as a second assignment method will do just that and as a result the classes will always be balanced. Both assignment methods are visually compared in Figure 2.3 and Table 2.2 below.

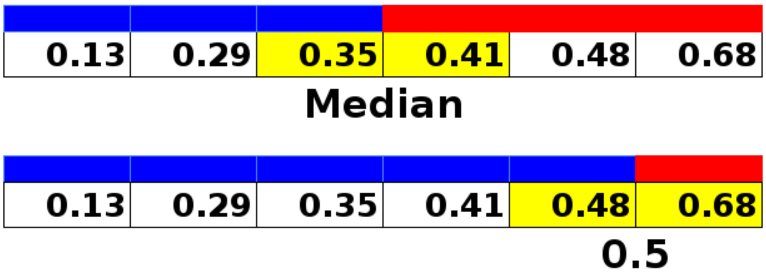


Figure 2.3: Assignments of different values to the two classes using two methods.

The results of different runs are obtained with leave-one-out validation where each of the 40 samples is predicted once, using the remaining 39 samples as training data. The 40 predictions' accuracies are then averaged which gave the following results:

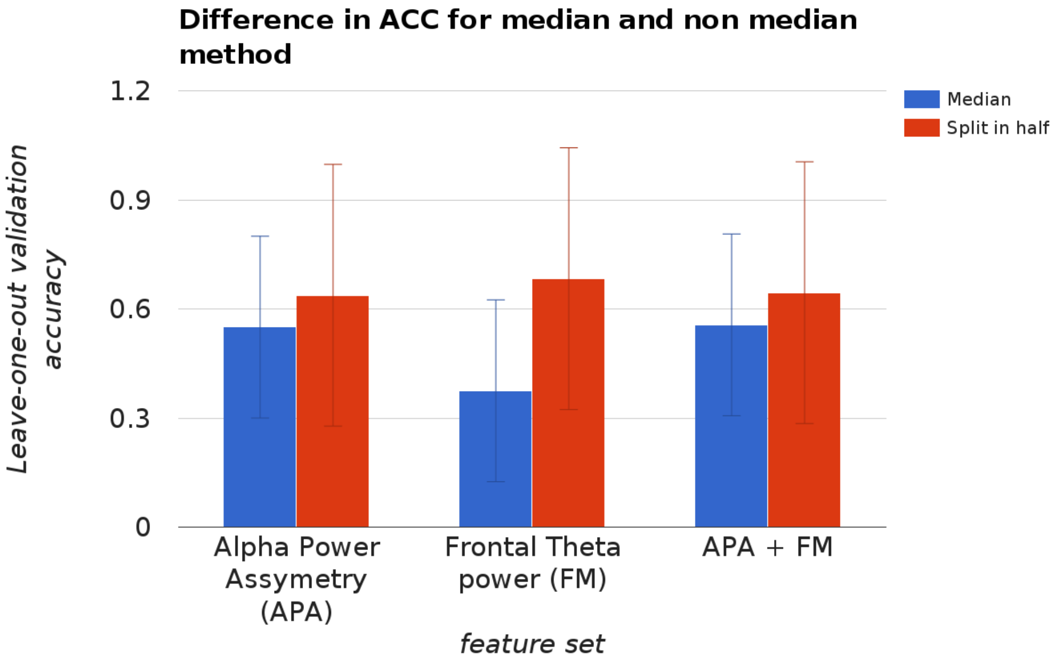


Figure 2.4: A comparison of the first results.

Features	**Median Method**	**split in half Method**
Alpha	0.55 ± 0.26	0.64 ± 0.37
FM	0.38 ± 0.25	0.68 ± 0.41
Alpha + FM	0.56 ± 0.28	0.64 ± 0.35

Table 2.2: Different accuracy values for the features sets and assignment methods (avg accuracy ± standard deviation).

Looking at the results it becomes clear that the average accuracy of the split in half method is higher than the median method. This is a result of assigning the classes with the median. For example with persons 6, some of the examples might end up in the low valence class even though person 6 might be feeling quite pleasant feelings during these EEG recordings, which confuses the model. Note that the standard deviation for the median method is lower than the split in half method, so in this perspective it is still possible that the results for the median method might be better.

2.3 CSP + LDA

Overall the results of the SVM explained above are reasonable for a first model, but higher accuracy is desired. Even though SVMs are capable of dealing with small datasets, it does not always provide the best results[18]. A frequently used model in the context of Brain Computer Interfaces is a combination of Common Spatial Patterns (CSP) and Linear Discriminant analysis (LDA) [19, 18, 20].

CSP is a spatial filtering technique that was originally designed to optimize the performance of motor imagery BCIs. CSP does so by creating linear combinations of the original EEG channels that maximize the variance for one class while simultaneously minimizing the variance of the other class [19]. One disadvantage of using CSP is that the default version can only distinguish between 2 classes, though one can easily aggregate multiple CSP models to create one-vs-one and one-vs-all models, similarly to the one-vs-one and one-vs-all SVMs.

LDA, on the other hand tries to find a projection of the data where the data is linearly separable.

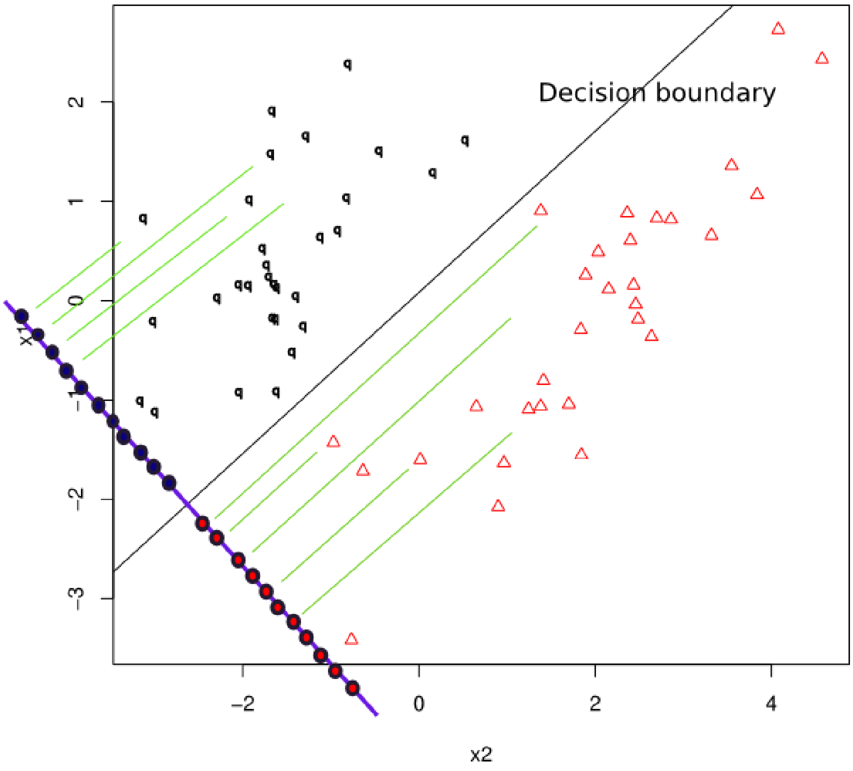


Figure 2.5: LDA finds a projection of the data where the separation of the data is clear.

To evaluate the model the dataset is first split in a train and testset (30/10), the testset is only used in the end to evaluate how well the model generalizes. During loading, the 32 original EEG channels are transformed to 16 CSP channels pairs, then the optimal number of channels pairs is determined used leave-one-out validation. This procedure thus starts with the outer channel pairs, performs leave one out validation and then uses the two outer channel pairs, etc. The set of channelpairs corresponding to the highest achieved validation accuracy is then used to predict the testset. Those results are shown in Figure 2.6.

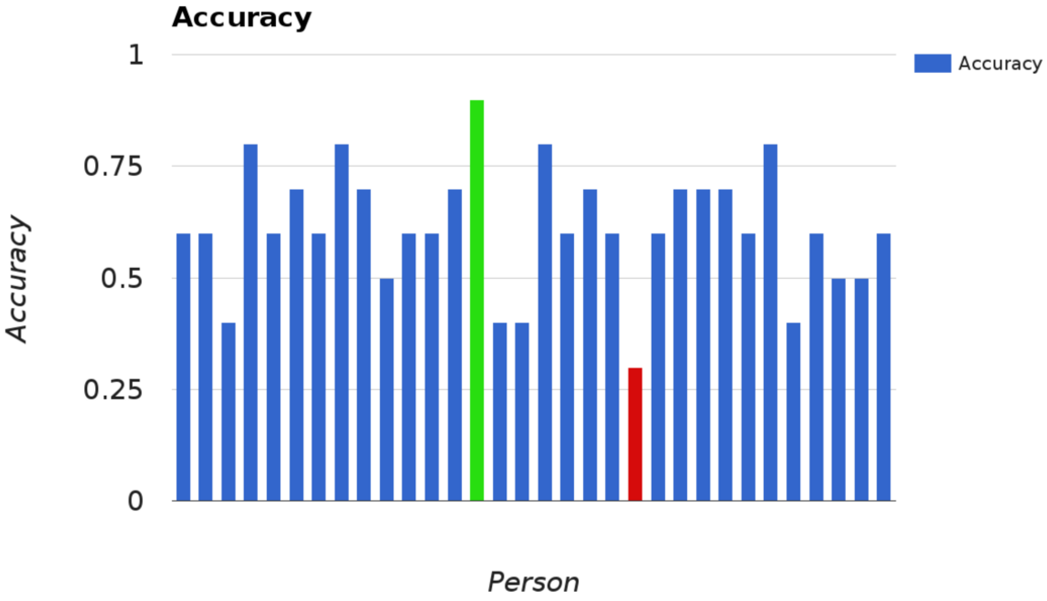


Figure 2.6: Accuracies for different persons, the highest (green) and lowest (red) accuracy are obtained by person 14 (90%) and 21 (30%) respectively.

From the results it is clear that this model actually works quite well for some persons while achieving lower accuracies for other persons. This variation in performance is typical for EEG and BCI related problems as EEG data is very person specific. Additionally one could plot the distribution of the original valence values, ranging between 0 and 1. Remember that these values were mapped to a label, where all valence values > 1 correspond to the high valence class while all valence values < 1 correspond to the lower valence class. Dividing these value in 8 bins and plot how many samples are present of each samples shows that for person 14 (highest accuracy) the greatest amount of samples is either very low or very high, and not a lot of mass is present near the classification boundary. For person 21 however, most of the samples are either just smaller of just greater than the classification boundary. With the absence of more extreme samples, it becomes hard for the classifier to achieve good performance.

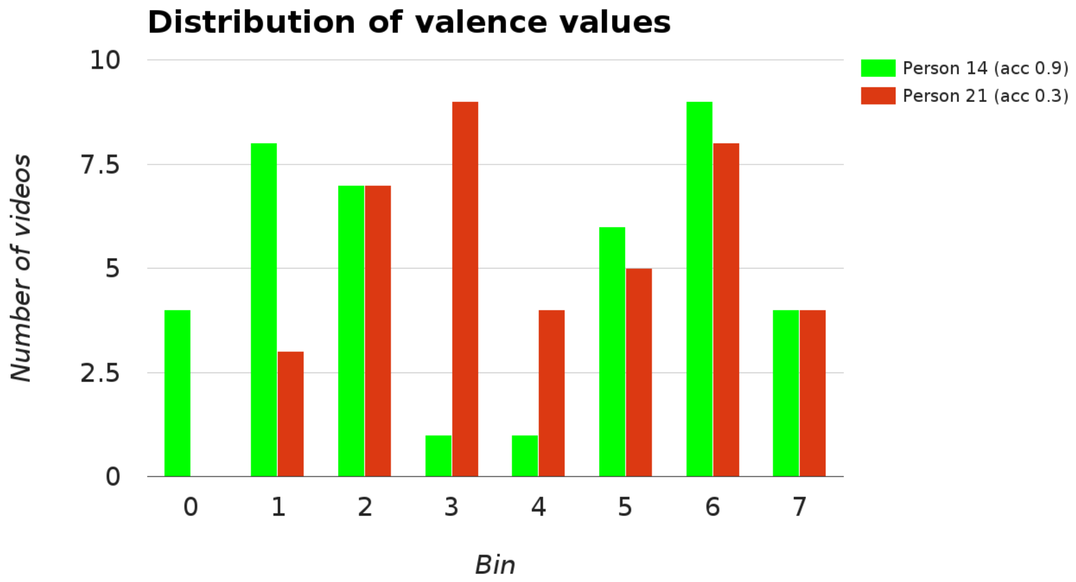
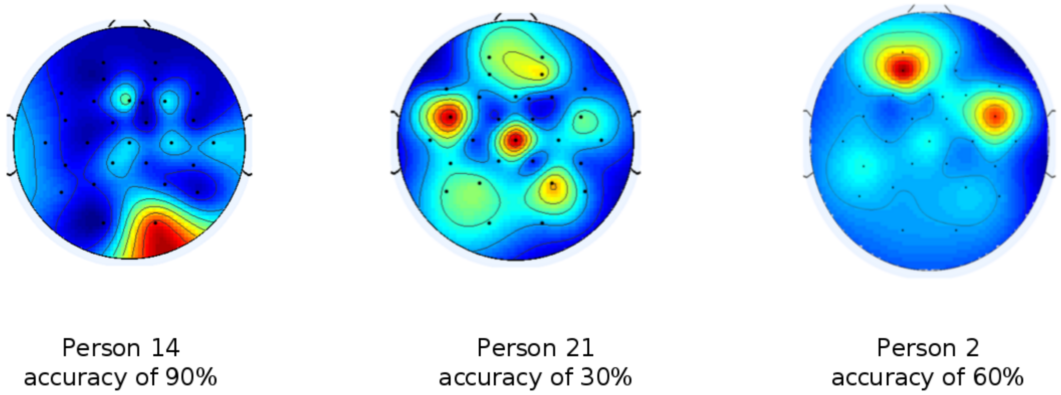


Figure 2.7: Distribution of valence values for the lowest and highest achieved test scores of the model.

Since CSP is applied it is also possible to look at the generated filter pairs to improve the understanding of the CSP algorithm. Figure 2.8 shows generated plots of the CSP values. The black dots represent given weights at the electrode locations the plot is then generated by interpolating the absolute weights of these points for the whole image. Blue colors correspond too less importance and red colors are the most important. Looking at the generated plots it become quite clear that the generated CSP filter is very person specific, CSP filter pairs of different persons look quite different.

Generally speaking, you would expect something similar to person 2 where the algorithm measures power components in the right and left parts of the scalp. This is somewhat the case for person 21, where the CSP gives high weights to different electrode locations. For person 14 however, the result is somewhat less desirable. Even though person 14 had the highest accuracy of 90%, the entire classification is based on a single electrode, which is an indication that the CSP is overfitting. This means that the performance might be great for the given test set, but its performance is likely to diminish quickly if one were to redo the experiment with the same classifier and a different set of samples.



Person 14
accuracy of 90%

Person 21
accuracy of 30%

Person 2
accuracy of 60%

Figure 2.8: CSP filters for different persons

3

Feature Selection methods

This section will first explain what feature selection means and why it is used, before giving a detailed overview of the different methods.

3.1 The need for feature selection

The need for feature selection is twofold: first by reducing the number of features, you can improve the accuracy, certainly when the training dataset is limited and there is a severe risk of overfitting. Second, reducing the number of features can increase performance. Additionally, in the context of research and this thesis, looking at which features are importance gives insight in the problem. For thesis it is important to find out which features are useful in the prediction of emotion.

3.2 Methods

There are several methods for the feature selection available, the following section will give a list of the most common used methods and their advantages / disadvantages.

3.2.1 Filter methods

These methods simple use a statistical test and filter out features that fail these test.

Removing features with low variance

This method simple removes all feature whose variance doesn't meet some threshold. Feature with have a low variance are less likely to be relevant, as there is very little variation in their value.

Univariate Feature Selection

The basic idea is to perform a statistical test to each feature individually. This is very simple, but will fail to recognise 'pairs' of important features; one feature may not be important on

label	**feature A**	**feature B**
Happy	+	+
Happy	-	-
Sad	-	+
Sad	+	-

Table 3.1: Some feature are not significant on its own, but a might be part of a combination of features.

its own, but might be a very good feature when combined with other features. Suppose the following example in Table **??**:

It is clear that feature A and B are very important when it comes to predicting whether or not a person is happy or sad. When both features have the same sign, the person is happy, otherwise he is not. This problem occurs in many simple selection methods.

3.2.2 Wrapper based methods

These methods use a training algorithm that guides them at finding the optimal subset of features. This methods is more recommended.

Recursive feature elimination

In this method you start by training a classifier with all the features and measuring the performance on a separate validation set. Once trained you look at the different coefficients that are assigned to each feature. Features with low coefficients are less important than features with high coefficient, so they can be removed. This process is repeated multiple times and every time a certain percentage of the lowest features is dropped until the validation score decreases significantly. This is done in L1 regression, you also have a less aggressive method , the L2 regression that lowers penalises the influence of features, but doesn't remove them, which strengthens its resistance against highly correlated features.

Randomized Sparse models

The main limitation of the above methods is that they will only select one feature out of a group of correlated features. A solution to solve this problem is to use randomization techniques, which randomly select features and look at the performance. combinations of features that give a high performance are then selected as being relevant. However this only works when the ground solution is sparse, meaning that only a small fraction of the features is relevant.

A sub group of randomized sparse models are the tree based estimators. For example, in a random forest features are randomly combined multiple times. The result is that each feature has an importance value, that indicates how important a certain feature is. Using the importance values, it is easy to remove certain features.

Coordinate Research

On coordinate descent the feature selection is formulated as a optimization problem and gradient ascent/descent is applied.

Bibliography

[1] H. Verschore, “A brain-computer interface combined with a language model: the requirements and benefits of a p300 speller,” afstudeerwerk, Ghent University, June 2012.

[2] D. O. Bos, “Eeg-based emotion recognition,” 2007.

[3] M.-K. Kim, M. Kim, E. Oh, and S.-P. Kim, “A review on the computational methods for emotional state estimation from the human eeg,” *Computational and Mathematical Methods in Medicine*, vol. 2013, no. 573734, p. 13, 2013.

[4] T. C. Technologies, *10/20 System Positioning manual.* Fortis Tower, 2012.

[5] unknown, “Electrode placement,” 2015.

[6] Y. Yang, S. Chevallier, J. Wiart, and I. Bloch, “Time-frequency optimization for discrimination between imagination of right and left hand movements based on two bipolar electroencephalography channels,” *EURASIP journal on Advances in Signal Processing*, vol. 2014, no. 38, 2014.

[7] K.-E. Ko, Hyun-Chang, and K.-B. Sim, “Emotion recognition using eeg signals with relative power values and bayesian network,” *International Journal of Control, Automation, and Systems*, 2009.

[8] Brainworks, “What are brainwaves?,” 2015.

[9] Y. Lio and O. Sourina, “Eeg databases for emotion recognition,” *International Conference on Cyberworlds*, 2013.

[10] Y. Lio, O. Sourina, and M. K. Nguyen, “Real-time eeg based human emotion recognition and visualization,” 2010.

[11] W. Zheng, J. Zhu, and B. Lu, “Identifying stable patterns over time for emotion recognition from EEG,” *CoRR*, vol. abs/1601.02197, 2016.

[12] R. N. Duan, J. Y. Zhu, and B. L. Lu, “Differential entropy feature for eeg-based emotion classification,” pp. 81–84, Nov 2013.

[13] Y.-P. Lin, Y.-H. Yang, and T.-P. Jung, “Fusion of electroencephalographic dynamics and musical,” 2014.

[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis ;using physiological signals,” *Affective Computing, IEEE Transactions on*, vol. 3, pp. 18–31, Jan 2012.

[15] P. LanG, M. Greenwald, M. Bradley, and A. Hamm, “Looking at pictures: affective, facial, visceral, and behavioral reactions.,” *Psychophysiology*, vol. 30, pp. 261–273, May 1993.

[16]

[17]

[18] F. Lee, R. Scherer, R. Leeb, C. Neuper, H. Bischof, and G. Pfurtscheller, “A comparative analysis of multi-class eeg classification for brain computer interface,” in *Proceedings of the 10th Computer Vision Winter Workshop*, pp. 195–204, 2005.

[19] A. Coone, “A study on different preprocessing and machine learning techniques for the detection of error-potentials in brain-computer interfaces,” afstudeerwerk, Ghent university, June 2011.

[20] G. Pfurtscheller, C. Neuper, C. Guger, W. Harkam, H. Ramoser, A. Schlögl, B. Obermaier, and M. Pregenzer, “Current trends in graz brain-computer interface (bci) research,” *IEEE Transactions on rehabilitation Engineering*, vol. 8, pp. 216–219, JUNE 2000.