A comparative study of physiological feature selection methods for emotion recognition

Andreas De Lille

Supervisors: Prof. dr. ir. Joni Dambre, Dr. ir. Pieter van Mierlo
Counsellor: Ir. Thibault Verhoeven

Master's dissertation submitted in order to obtain the academic degree of
Master of Science in Computer Science Engineering

Department of Electronics and Information Systems
Chair: Prof. dr. ir. Rik Van de Walle
Faculty of Engineering and Architecture
Academic year 2015-2016

UNIVERSITEIT GENT

Contents

1

Introduction

This chapters introduces the masterthesis. It starts by explaining what a brain computer interface is and how it works. After that, emotion recognition is explained and basic concepts of machine learning are introduced. The last section of this chapter covers the goal of the thesis.

1.1 Brain computer interfaces

A Brain Computer Interface (BCI), creates a direct neural link between the brain and the computer[**?**], that tries to recognize patterns and based on the extracted information, performs actions. A BCI removes the 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 analyse brain activity, the most convenient method is Electroencephalography (EEG), since it is a non-invasive method. Non-invasive methods, in contrast to invasive methods require no surgery; in the case of EEG, they simply measure electrical activity using electrodes placed on the scalp.

Electrical activity in the 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[**?**]. 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 neighbouring neurons. When this reaction occurs simultaneously in a lot of neurons, the change in electrical potential becomes significant, making it visible to the EEG surface electrodes. EEG can thus only capture synchronized activity of many, many neurons, which explains its low spatial resolution capabilities.

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 of noise to the signal,

which explains why EEG signals are very noisy signal from nature. Noise removal techniques are therefor advised. Note that even though EEG data contains a lot of noise and has a low spatial resolution, it still provides significant insight into the electrical activity of the cortex while offering excellent temporal resolution[**?**].

Note that EEG only records electrical activity of the brain, other methods like magnetoencephalography (MEG) use magnetic fields to measure brain activity. 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. As a result, this method was not explored during this thesis.

EEG measures electrical activity with electrodes that are placed on the scalp. 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[**?**]. 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 a number that determines the 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. Figure **??** shows a picture of a 23 channel 10/20 system. Note that the 10/20 system does not require a fixed number of channels, some experiments may use a different set of channels, but they all follow the same naming convention.

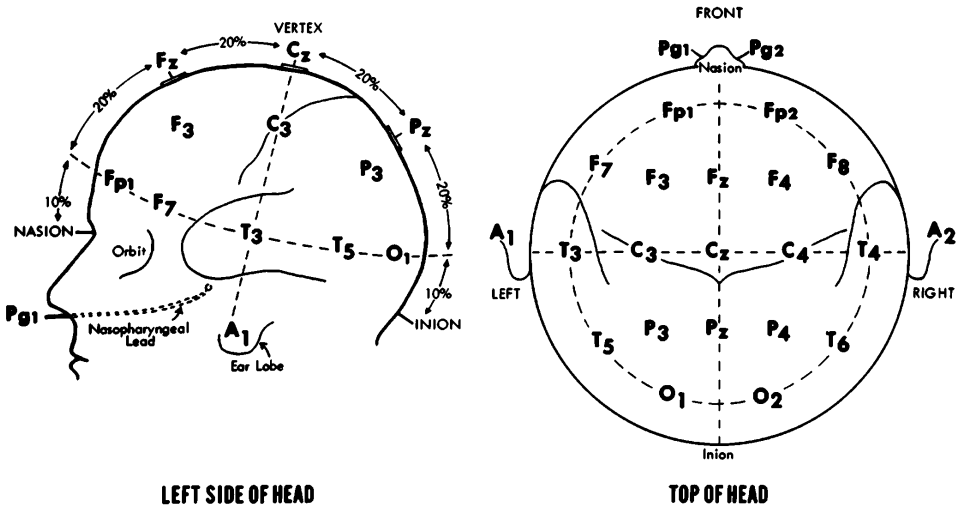

Figure 1.1: The electrode placement of a 23 channel system[**?**].

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, on the otherhand, is obtained by subtracting two monopolar EEG signals, which improves Sound to Noise ratio (SNR) by removing shared artifacts[**?**].

In the frequency domain, brain waves are usually split up into different bands[**? ?**], 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, occurs during dreaming.

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

1.1.2 Person specific classification versus cross-subject classification

There exists two types of BCI applications. The first type is a person specific BCI, where the BCI interface is calibrated for a single subject. The second type is a general BCI interface that works 'cross-subject', meaning that it should be able to work for different persons. It is much harder to achieve good results for a cross-subject BCI, as EEG data is very personal from nature[**?**]. While transfer learning has provided good results in imaginary motion recognition in the past, little research has been done for transfer learning in the context of emotion recognition. Person specific classifiers are often used, due to the fact that finding person independent EEG features is still an ongoing topic of research[**?**].

1.2 Emotion recognition

Psychology makes a clear distinction between physiological behavior and the conscious experience of an emotion, called expression[**?**]. 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. This makes emotion recognition based on physiological signals more robust to social masking[**?**], the process where an individual masks or hides their emotions to conform to social pressures. To really know one's emotions, it seems, one has to research the physiological aspect of the emotion.

1.2.1 Valence/Arousal classification model for emotion

Before emotions can be recognized, a classification model is needed. A simple way of achieving this is using several discrete emotions, e.g. anger, joy, sad and pleasure. A more convienent model to classify emotions is the bipolar arousal-valence model[**? ?**], that places emotions in a two dimensional space. The main advantage of using a continuous 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 **??** shows the mapping of different emotions for this model.

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. Even though arousal and valence describe emotions quite well, a third dimension, dominance, can also be added. The third dimension, 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.

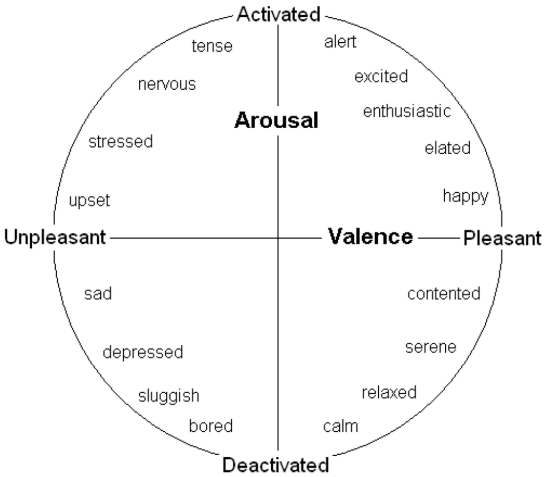

Figure 1.2: The arousal - valence model maps emotions in a two dimensional plane.

1.2.2 Possible applications for emotion recognition

Emotion recognition has many different applications, e.g. as an improvement for the P300 speller or marketing analysis. The P300 speller is a very well-known, academic application of BCI and an active topic of research. It uses EEG data to enable patients with a locked in syndrome to communicate[**?**]. The basic version uses a six by six grid of characters, each row and column is flashed in a random order while the subject silently counts the number of flashes of a certain character, as shown in figure **??**. This procedure, where a train of stimuli with some infrequent occurring target stimuli is applied, is called the oddball paradigm[**?**]. It is known that this technique triggers an increase in the potential difference in the EEG around the parietal lobe. When a potential difference in the brain occurs as a reaction to an event, it is referred to as en event related potential (ERP) . The P300 ERP occurs roughly 300 milliseconds after the stimuli is flashed, hence its name[**?**]. The presence or absence of the P300 waveform is used by the P300 speller to determine what character the subject was focusing on, which basically allows the subject to spell text.

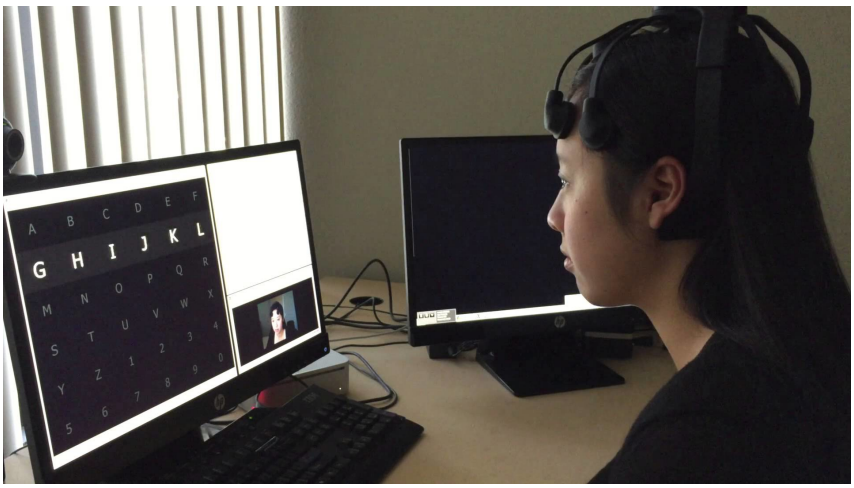

Figure 1.3: Different parts of the P300 speller, found at [**?**].

Research with visual stimuli on healthy subjects, has shown that emotion has an effect on the auditory P300 wave[**?**]. Both the P300 peak amplitude and area was highest when viewing neutral pictures and descended further, in decreasing order, for sadness, anger and pleasure. The latency of the P300 ERP speller was shortest or neutrality and in increasing order longer for pleasure, anger and sadness. It is expected that a visually triggered P300 wave, will also be influenced by emotion. Having a good emotion recognition system, might therefore improve the detection of P300 waves. Additionally knowing a subject's emotional state can help detecting when a subject gets frustrated, e.g. because of mistakes he makes.

An improvement in performance is not the only advantage an emotionally aware P300 speller has. Contrary to what subjects might think, the P300 speller is unable to read the mind and know what a person is thinking about[**?**]. The P300 speller provides no more than a means of communication that the subject can use. Should he chose to ignore the instructions and focus his attention elsewhere, then the recordings become useless. Nevertheless, ethical questions often remain unanswered. Knowing how the subject feels, can help him communicate more humane on one hand, while providing more insight for ethical issues, on the other hand, e.g. "How does the

subject think about the P300 speller recording and analyzing his brain activity?". Information about the subject's emotional state can help answer some of these ethical questions. Integrating the results from this thesis with the P300 speller, is an opportunity for future research.

Another application for emotion recognition is in the field of marketing and customer satisfaction research. Discovering how a persons feels about a product is often tricky. Questionnaires is one way to go, but they might contain a lot of noise. Being able to 'read' the emotion straight from a subject's mind, is expected to give more accurate results as it avoids any form of social masking.

1.3 Machine learning

Machine learning is the missing link between the EEG data the emotion recognition. Machine learning is a very broad domain, as a result, this discussion will be limited to an introduction of the basic machine learning concepts with the focus on the application of machine learning and machine learning techniques.

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.

To further explain how machine learning works, have a look at the following example. 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. The total area is a characteristic of the house that helps us in predicting the price. In the context of machine learning, the characteristic 'total area', will be called a feature as the asking price of a house is correlated to the total area.

One possible way of predicting the asking price of a house is machine learning. Machine learning works in several steps, first you need to train your machine learning algorithm with a list of asking prices and the corresponding area of the house. This process is called training or fitting and gives the machine learning component an idea to what the corresponding price is for an area. The outcome might be a coefficient, suppose one square meter is worth 1000 Euro, than the predicted asking price will be 1000 x the total area.

Even though, this might already give some reasonable results, the algorithm will probably not be accurate enough for real life usage. This is due to the fact that the area of the house is only one feature that determines the price. Other features, like the number of bedrooms or the location of the house, were not taken into consideration. Adding additional features, gives more insight into the data, e.g. a house with 5 bedrooms is more expensive than a house with only 3 bedrooms. Having more features is thus likely to improve the performance of the machine learning algorithm.

Machine learning algorithms are responsible for finding the relation between features and the predicted value. There are many of these 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 (more or less) a continuous value. Machine learning problems that require

the output of a continuous value, are called regression problems. In the OCR example from above, where characters are recognized in a picture is a classification problem, as the outcome is a character out of a limit set of different characters.

Another way to group algorithms is on their training data. In the asking price examples above, the training data consists of labelled results; the correct asking price is given for each area. This type of machine learning, where the correct outputs for the train data are given, is referred to as supervised machine learning. The alternative 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, known as semi supervised learning, is also possible. Imagine a dataset with 5000 webpages that need to be grouped into 10 distinct categories, e.g. science, nature, cooking, ... Only 100 of the 5000 pages in the train set are labelled. An approach to solve this problem could be to first cluster the pages in similar groups using unsupervised learning. As soon as a group contains a single labelled page, all pages in the group can be labelled accordingly, as clustering returns groups of similar samples. Semi supervised learning has the advantage that one can also use unlabelled data, which is often easier and cheaper to obtain, unlike labelled data which is usually quite rare; if there was a fast and easy way to label the data then there would not be a need for machine learning.

Machine learning can be used for emotion recognition to find patterns in features extracted from physiological signals. The output of the machine learning algorithm is a prediction of the subject's emotional state. The general process of machine learning is as follows, the process starts with gathering EEG data, from which features are extracted. These features are then fed to a machine learning algorithm, which outputs a prediction.

To recognise emotion in the brain, features need to be extracted from the physiological signals. In this thesis two categories of physiological features are observed. non-EEG features and EEG features. Non-EEG features are physiological signals like heart rate, skin conductivity, respiration rate, ... These features are more convenient to obtain as they do not require one to mount an EEG cap to a subjects scalp.

EEG features, on the other hand, are features extracted from the EEG measurements of the subjects. Unfortunately, the literature does not fully agree on a specific set of features nor does it agree on what channels and/or waveband are most important for emotion recognition. This problem is addressed in this thesis, by ranking a large set of features using different features selection methods. However the literature does agree on certain things; the right hemisphere of a subject is generally speaking, more active during negative emotions than the left hemisphere, which is in turn more active during positive emotions[**? ? ?**]. The features are discussed in more depth in **??**.

1.3.1 Over and underfitting / high bias and high variance

Over and underfitting is a common problem in many machine learning projects. An algorithm that 'overfits the data' is able to recognise seen sample points very well. However when the algorithm is tested on unseen points, the performance might be a lot lower than expected.

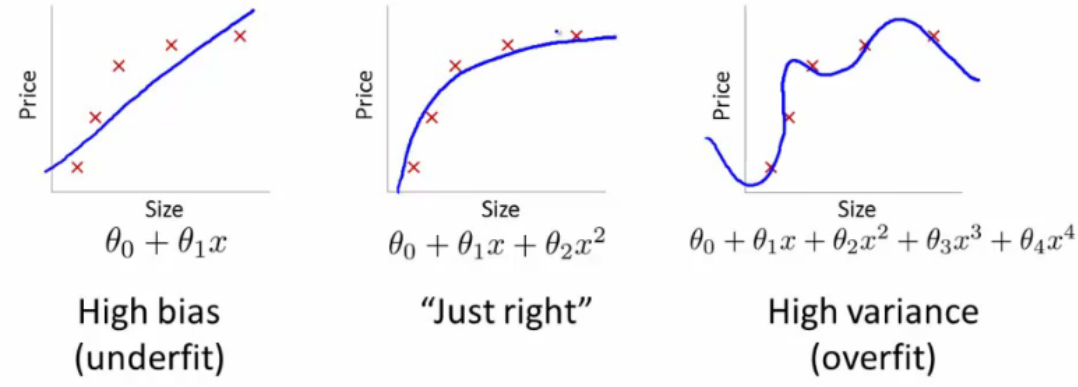

Figure 1.4: Overfitting versus underfitting[**?**].

Suppose the example in Figure **??**, 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 corresponds to the most logical 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. This is known as a high bias problem. A high bias problem has a high training error, as the function is not able to fit the points sufficiently, this is visible in Figure **??**.

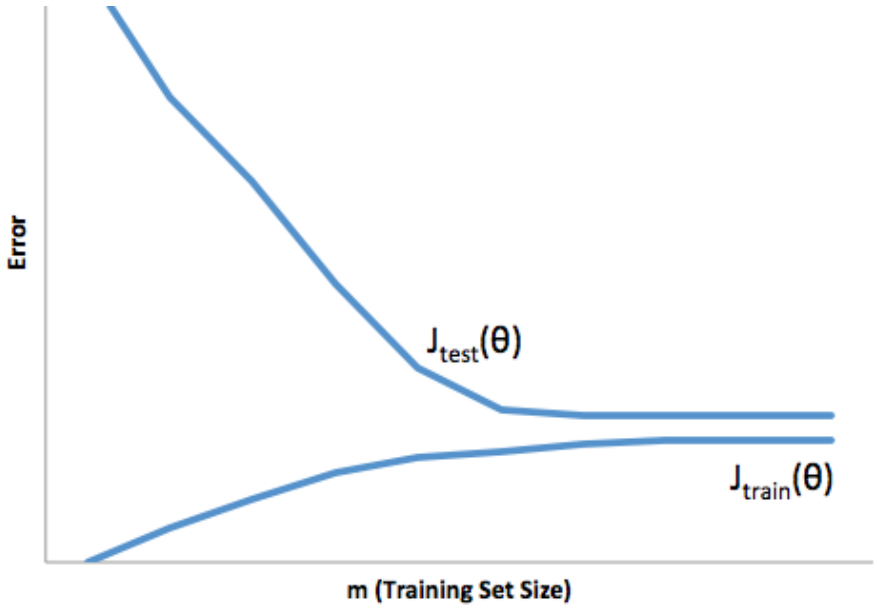

Figure 1.5: A high bias function is not complex enough to approach the generator function closely.

The function on the right corresponds to an overfit; the function fits or 'goes through' each point exactly, but one can see that the behaviour of the hypothesis function in between data points is

not what one would expect. This problem is known as a high variance problem. A high variance problem occurs when the train error is close to zero, so the algorithm fits the points well, but the test error is quite dramatic, which means that the algorithm will perform very badly when it gets new points.

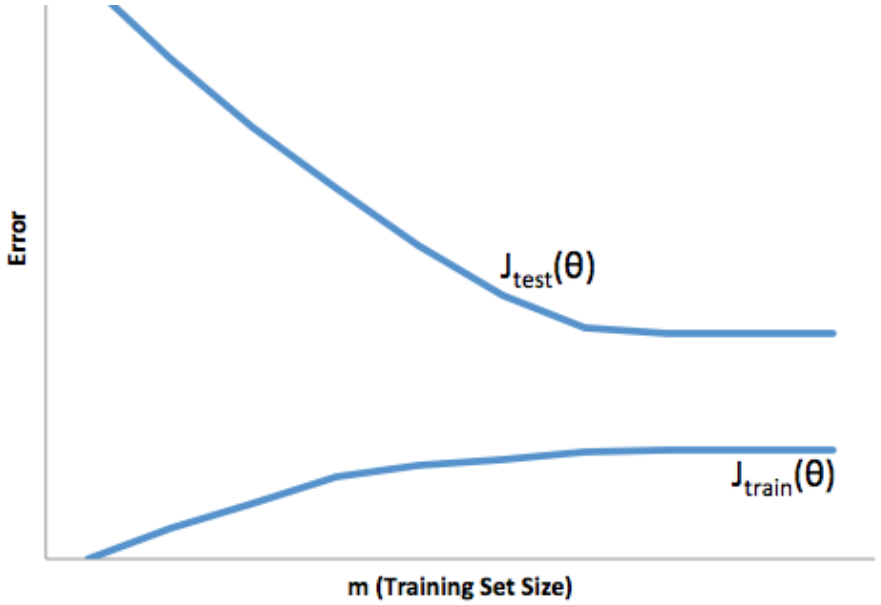

Figure 1.6: A High variance function is too complex and fits the data point too closely.

Another way to explain the bias variance trade-off is by an example. Suppose you have a dart board, as shown in Figure **??**. 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, in other words, 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 a high variance problem; the solution is right in front of you, but you are not able to reach it precisely.

Figure 1.7: The bias variance explained using the dartboard example found at [**?**]

1.4 Goal of the thesis

The first goal is finding relevant features for emotion recognition in a person specific setting. This is already quite challenging as there are fuzzy boundaries and individual variation of emotion[**?**]. To do so, the output of different feature selection methods is compared. In a successful scenario, good features are found that, once given to a machine learning algorithm, can accurately predict the emotions of one person. In this stage some attention will also be spend on comparing non-EEG and EEG features to see which feature set contains the most information.

The second goal is finding features for emotion recognition in a cross-person 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 comparison for non-EEG and EEG features will also be done here. Emotion recognition is harder in a cross-subject setting, since physiological signals are very personal[**?**].

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. Using fewer features, does not only limit the risk of overfitting, it might speed up the algorithm and preparation time. Mounting EEG electrodes is a time consuming activity, using fewer electrodes limits this time.

Another point to note is that even though a simple limited subset of features might solve the overfitting problem, it will likely result in a performance drop as optimal features might have been left out. This problem is even more severe in a cross person setting, 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 work 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

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
[**?**]	alpha/beta ratio, Fpz beta and alpha band power
[**?**]	PSD, RCAU, DCAU, DASM, RASM, DE

limited and person specific set of good features is selected. This allows the machine learning to use good features, while keeping the set of features limited in size. Another solution is to use dimensionality reduction, to project the feature space to a lower dimension, but this has the disadvantage that bad features still have influence and thus might confuse the classifier.

Additionally, a lot of different features are reported in the literature, as you can see in Table **??**. This thesis tries to overcome this problem, by comparing a large set of features with different feature selection methods to the features reported in literature.

1.4.1 Dataset

One of the most used datasets in the context of emotion recognition is the Dataset for Emotion Analysis using Physiological Signals (DEAP) dataset[**?**]. This dataset consists of several parts, the first part is a rating of 120 music videos by 14 - 16 persons. Each video is rated for valence, arousal and dominance on a scale ranging from 1 to 9. This part of the dataset is not used during this thesis, because it contains no EEG recordings.

The next part of the dataset is the physiological experiment that contains emotional reactions of 32 subjects. The emotional reactions were triggered using music video excerpts; each subject watched 40 one-minute videos, while several physiological signals were recorded. These physiological signals consist of 32 channel 512Hz EEH and peripheral physiological signals. More concretely, this dataset contains following signals:

There also exists a preprocessed version of the physiological experiment database, where the EEG recordings were downsampled to 128Hz and noise and EOG artifact removal was performed. This dataset is used during this thesis.

Additionally facial video for 22 of the 32 subjects was recorded, so research in facial expressions is also possible with this dataset. These videos are also rated on 4 scales: arousal, valence, dominance and liking. 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. However strong correlations were observed[**?**]. 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[**?**]. SAM visualizes the valence, arousal and dominance scales with pictures, each picture corresponds to

Table 1.2: The available signals in the DEAP dataset

Channel	Name	Category	Channel	Name	Category
1	Fp1	EEG	**21**	F8	EEG
2	AF3	EEG	**22**	FC6	EEG
3	F3	EEG	**23**	FC2	EEG
4	F7	EEG	**24**	Cz	EEG
5	FC5	EEG	**25**	C4	EEG
6	FC1	EEG	**26**	T8	EEG
7	C3	EEG	**27**	CP6	EEG
8	T7	EEG	**28**	CP2	EEG
9	CP5	EEG	**29**	P4	EEG
10	CP1	EEG	**30**	P8	EEG
11	P3	EEG	**31**	PO4	EEG
12	P7	EEG	**32**	O2	EEG
13	PO3	EEG	**33**	hEOG	non-EEG
14	O1	EEG	**34**	vEOG	non-EEG
15	Oz	EEG	**35**	zEMG	non-EEG
16	Pz	EEG	**36**	tEMG	non-EEG
17	Fp2	EEG	**37**	GSR	non-EEG
18	AF4	EEG	**38**	respiration belt	non-EEG
19	Fz	EEG	**39**	plethysmograph	non-EEG
20	F4	EEG	**40**	temperature	non-EEG

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. In this thesis, a preprocessing step scaled and translated these values to ensure they range between 0 and 1, a more convenient interval.

The used SAM figures are shown in Figure **??**. 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.

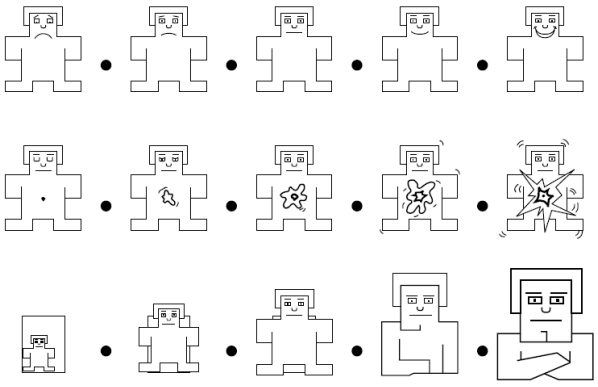

Figure 1.8: The images used for the SAM[**?**].