A Comparative Study of Physiological Feature Selection Methods for Emotion Recognition

Andreas De Lille
Promotors: Prof. dr. ir. Joni Dambre and dr. ir. Pieter Van Mierlo
Counsellor: ir. Thibault Verhoeven

Abstract—An emerging topic of research is emotion recognition based on physiological signals and machine learning. Emotion recognition is the process of recognizing a person's emotional state. In this work the emotion recognition was done using a combination of physiological signals and machine learning. The general flow of this approach is to record physiological signals from a person, extract features and feed them to a machine learning algorithm. This algorithm will then predict the user's emotional state. Even though a lot of research has been done, there is no agreement on what features are important. This work tries to overcome this problem by comparing a wide range of features with several feature selection methods.

Index Terms—Emotion recognition, physiological signals, machine learning, feature selection methods

1 Introduction

Emotion recognition is the process of recognizing a person's emotion. Observing and recognizing emotion can be done in several ways. Psychology makes a clear distinction between physiological behaviour and a person's expression of emotion[1]. The expression is often prone to social masking[2], the process of hiding emotion to conform to social standards and ideas, making it less reliable. The physiological behaviour on the other hand is much harder to control, making it more reliable. This work will thus focus on emotion recognition based on physiological signals.

In the next section, the classification of emotions will be explained. Before an introduction of physiological signals is given in Section 3. What follows in Section 4 is an overview of the used features. The problem statement and goal of this work are given in 5. Section 6 gives the used approach and results. In Section 7 the conclusion is given.

2 Classification of emotions

Before emotion can be recognized, different emotions need to be defined. One way to do this is to use several distinct emotions, e.g. anger, joy, sad and pleasure. The advantage of this approach is that all emotions have a clear label. The disadvantage is that this model is often not complex enough to represent the whole emotion space. To solve this problem, the bipolar valence-arousal model was introduced[1, 3]. This model puts each emotion in a two dimensional space. The first dimension indicates how active a person is feeling. The next dimension is valence, which indicates how pleasant or unpleasant the emotion is perceived.

The valence-arousal model has the advantage that an emotion can be defined, without the explicit need for a label. All discrete emotions are mapped to the valence-arousal space. For example, excitement corresponds to an active feeling with a pleasant experience, meaning that it will be in the high valence, high arousal quadrant of the space. A depressed feeling, on the other hand, will have a low valence and low arousal. As a result depressed is mapped in the low valence, low arousal quadrant. The mapping of other emotions can be done similarly and is shown in Figure 1.

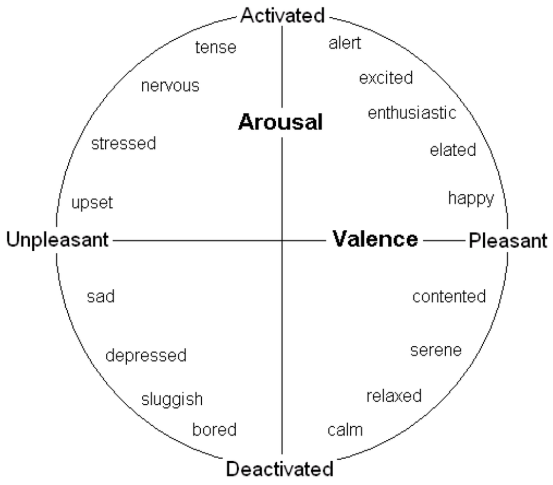

Fig. 1. The arousal - valence model maps emotions in a two dimensional plane.[4]

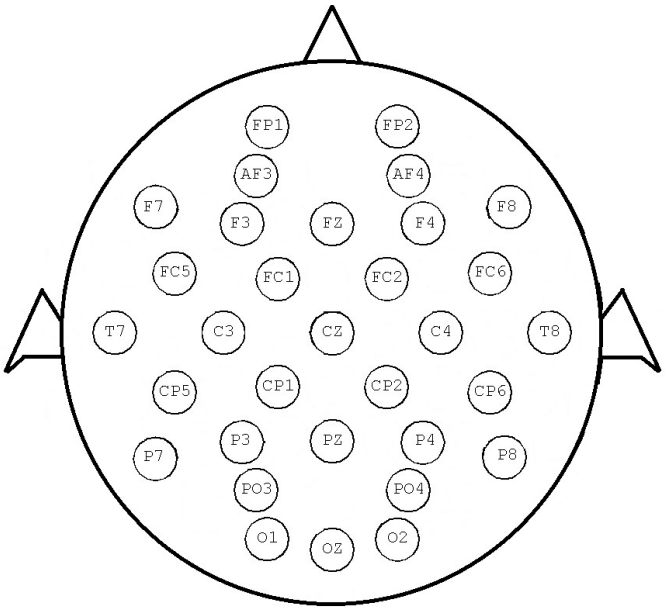

Fig. 2. Placement of the 32 electrodes in this work.[6]

3 Physiological signals

In case machine learning is used for emotion recognition, some physiological signals are taken as input. The machine learning will then output a valence or arousal score. In short, machine learning can be defined as an input output model that predicts output values for different samples based on the inputs. The inputs are features of the input samples, e.g. the frequency or amplitude of a signal.

To do emotion recognition with machine learning, good features are required. This work focusses on physiological signals from which two groups of features can be extracted. The first group contains the peripheral signals, a.o. heart rate, blood pressure, respiration rate, perspiration, etc. The second group of features originate from the brain. These signals are recordings of brain activity using electroencephalography (EEG). EEG is a technique that measures electrical activity of the brain, by placing electrodes on the scalp[1]. EEG is very noisy by nature as the signal is distorted by the bone between the cortex and the electrodes. Still, it provides significant insight in the brain activity[5]. All electrodes are placed according to the 10/20 system, that labels each location. The locations and the corresponding channel names used in this work are visible in Figure 2.

EEG measures electrical activity at each channel. Each measurement can be split in different frequency bands, with medical relevance[7, 8]. The frequency bands are:

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

4 Features

In this work the Dataset for Emotion Analysis using Physiological Signals (DEAP) was used. This dataset consists of recordings of a physiological experiment[9]. This experiment recorded emotional reactions of 32 subjects. Each subject watched 40 one-minute video excerpts to trigger emotions, while physiological signals were recorded. These physiological signals consist of 32-channel, 512Hz EEG signals combined with the following peripheral signals:

- galvanic skin response (GSR), which measures perspiration
- respiration belt, which measures the respiration rate

- plethysmograph, which measures the blood pressure
- skin temperature

For the peripheral signals, statistical values of each channel are used. These statistical values are minimum, maximum, variation, standard deviation, average and median of each channel.

EEG features use a different approach, here the power spectral density (PSD) of each EEG signal is calculated. The PSD gives the distribution of the signal's energy in the frequency domain. Another power feature is the differential entropy (DE), it is proven that the differential entropy of a certain band is equivalent to the logarithmic power spectral density for a fixed length EEG sequence[10].

The most used features for valence classification are asymmetry features that measure asymmetry between two channels[5]. This can be done in four ways. The first way is the differential asymmetry (DASM) which is defined as:

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

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

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

Instead of looking at the asymmetry between left and right, one can also compare the frontal power with the posterior power[5]. This is known as the caudality. The differential caudality (DCAU) and rational caudality (RCAU) are defined as:

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

Another category of features are the power fractions, which give the power distributions for each channel[1]. They are defined as:

$$frac_{band,channel} = \frac{power_{band,channel}}{power_{total,channel}}$$

Often the ratio of Alpha/beta power is also used for classification of arousal[11].

5 Problem statement

Looking at different features in Table 1, one can see that the features quickly add up. Having a total of 894 features will increase the risk for overfitting significantly[12]. Additionally, not all 894 features are important, and there is no agreement on the most important features in literature.

TABLE 1
An overview of the different features that were compared in this thesis.

Name	Type	Channels	Freq bands	Total
PSD	EEG	32	6	192
DE	EEG	32	6	192
DASM	EEG	13	6	78
RASM	EEG	13	6	78
DCAU	EEG	11	6	66
RCAU	EEG	11	6	66
Frac	EEG	32	5	160
Alpha / Beta	EEG	32	1	32
EEG Total				864
Name	**Type**	**Total**		
HR	non-EEG	6		
Plethysmograph	non-EEG	6		
GSR	non-EEG	6		
ST	non-EEG	6		
RSP	non-EEG	6		
non-EEG Total		30		
Overall Total	**894**			

The goal of this work is twofold, first important features are needed for person specific emotion recognition. Second, features are needed for a cross-subject setting, where the system is trained and tested on different persons. Working in a cross-subject setting is more challenging, because physiological signals are personal by nature[9].

6 Used approach and results

Several feature selection methods were compared. A feature selection method is a method that takes as input a large set of features, and returns a (smaller) set of important features. The feature selection methods are grouped into three categories: filter, wrapper and embedded.

Filter methods are less complex methods that simply use a statistical test to rank the features. An example would be to use the Pearson correlation between a feature and the valence. Features with low of zero importance would then be filtered. This method is fast and simple, but is not capable to find groups of relevant features, as each feature is reviewed on its own.

Wrapper methods are more advanced. They use an arbitrary machine learning technique and look at the assigned weights. Features with large weights

have more influence on the output than features with low weights. Note that absolute values are used; large positive and negative values have the same amount of influence on the result.

The last category of feature selection methods are the embedded methods. Like wrapper methods, embedded methods also use an underlying machine learning technique, but with build-in feature selection functionality. Random forest are a good example of this category. Random forest use multiple decision trees that each have a random subset of the samples and features[12]. Building a decision tree starts with all samples in one node. This node will therefore be impure, meaning that it has samples of multiple classes. In the next step, the data will be split iteratively. In each iteration, a features is picked at random and the data is split according to the different values of features. This split is done in order to lower the impurity of the child nodes. After several iterations, the node will eventually be pure. A pure node contains only samples of one class. Each split corresponds to a drop in impurity, cause by evaluating the chosen features. By averaging the drop in impurity over all trees, one can estimate the importance of the different features quite well. The advantage is that this method is capable of finding groups of relevant features. When a group of random features is selected at different nodes in a decision tree, the impurity will drop significantly. This will result in a higher feature ranking.

The following approach was used, first the DEAP dataset is split in a train and test set (30 and 10 samples respectively). The test set is kept separate, while the random forest feature selection is performed on the train set. This step returns the top 30 features. 30 was selected after observing that 30 features was large enough to contain enough relevant features. With these 30 features a model is build by iteratively repeating the following steps:

1) add a new feature to the feature set
2) determine the cross validation error and standard deviation (std)
3) if the validation error and standard deviation are better than the previous best, the feature is kept. Otherwise the feature is neglected. This was done to increase the stability of the method and was inspired by literature [12].

The model that is being build is an SVM with an Radial basis functions (RBF) kernel, sometimes referred to as a Gaussian SVM. This model was chosen because it has proven itself in multiple emotion recognition studies[10, 13, 14, 15]. SVMs look for a separation boundary between two classes, and thus only look at points close to that boundary. This gives this method an advantage, as the dataset only contains 40 samples for each person, only 30 of them are available for training. Another advantage is that SVMs are capable of handling large features sets[16, 17]. Both advantages concur with this work's problem statement.

For each person a model is build and tested on the test set, the resulting performances are averaged over all persons giving a performance of 70 and 73.75 % with std 13.9 and 13.1 % for arousal and valence respectively.

To review which features were selected, groups of similar EEG features were created. These groups are:

- Power features: the PSD and DE features
- Asymmetry features: DASM, RASM, DCAU, RCAU
- Fraction features: the fractions

The non-EEG features are:

- heart rate
- GSR
- respiration rate
- blood pressure
- skin temperature

The distribution of the selected features is given in Figure 3 for arousal and Figure 4 for valence, the legend is given in Figure 5.

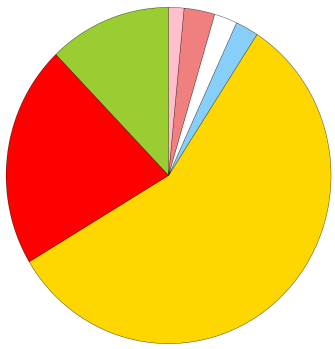

Fig. 3. Distribution of the selected features for arousal classification.

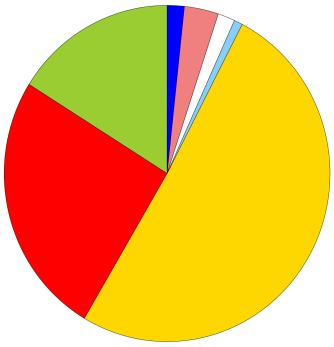

Fig. 4. Distribution of the selected features for valence classification.

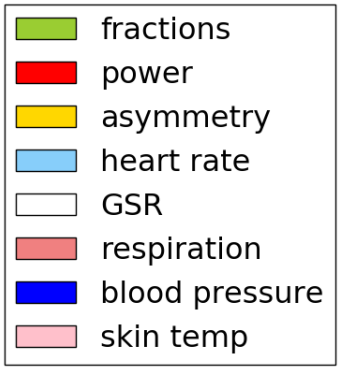

Fig. 5. Legend for Figure 3 and 4.

It is clear that for both arousal and valence, the asymmetry features are the most important. The second most important category are the power features. The non-EEG features seem to have very little influence on the result. To verify that the non-EEG feature are worse than the EEG features, the setup was run two additional times. One time with only EEG features and one time with only non-EEG features. The results are visible in Figure 6 and Figure 7 for arousal and valence respectively.

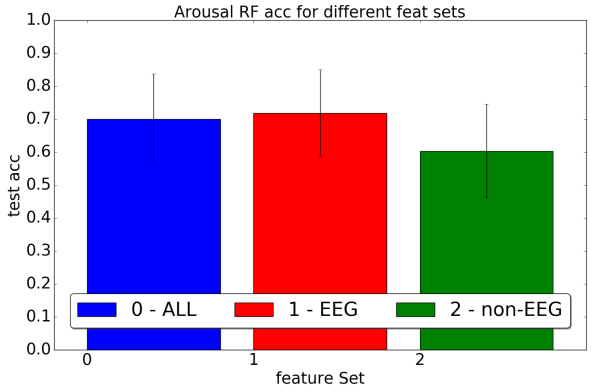

Fig. 6. The performance of arousal prediction for all, EEG and non-EEG features. The reported accuracies are test accuracies averaged over all persons, with their standard deviation.

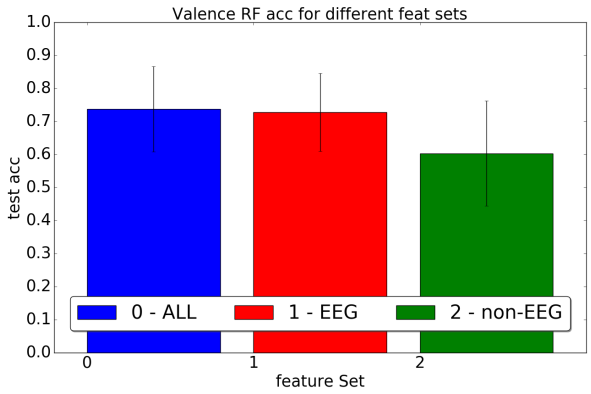

Fig. 7. The performance of valence prediction for all, EEG and non-EEG features. The reported accuracies are test accuracies averaged over all persons, with their standard deviation.

The results were compared with a two-sided p-test which does confirm that the non-EEG features have a lower performance. The P-values are shown in Table 2.

TABLE 2
P-values for the comparison of the performance of different feature sets.

	all / EEG	all / non-EEG	EEG / non-EEG
Arousal	0.4386	$5.891 * 10^{-7}$	$1.201 * 10^{-4}$
Valence	0.6817	$1.993 * 10^{-9}$	$1.763 * 10^{-6}$

For the stability, the Jaccard index, an index that measures the similarity between two sets, was calculated. A Jaccard index of zero corresponds to two totally different sets, while a Jaccard index of one

corresponds to two identical sets. For the Random forest the average Jaccard index over all persons was 0.743 with a standard deviation of 0.24 for valence and 0.791 with a standard deviation of 0.244 for arousal. The large standard deviation indicates that even though the RF method is quite stable, results vary from person to person.

For the cross-subject setting, performance was much lower than the person specific setting. This was expected, as physiological signals are very personal by nature **??**. The test accuracy was 64% for arousal and 55% for valence. The drop in performance, caused by the transition from person specific to a cross-subject setting was larger for valence than arousal. This indicates that physiological reactions to a change in arousal are more common between persons than a change in valence.

The selected features are similar to the person-specific case. Asymmetry EEG features are selected most of the time and non-EEG features were rarely selected. However when comparing performance of three feature sets containing all, EEG and non-EEG features, the non-EEG feature set scored higher than in a person specific setting.

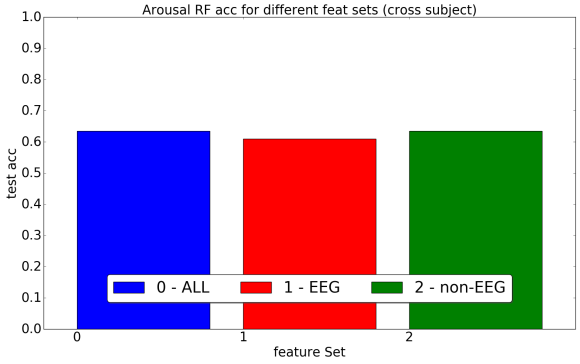

Fig. 8. The performance of arousal prediction for all, EEG and non-EEG features.

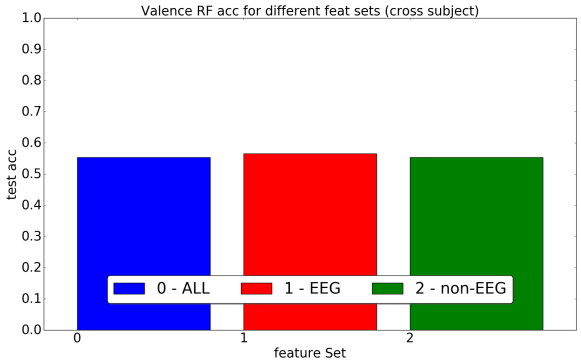

Fig. 9. The performance of valence prediction for all, EEG and non-EEG features.

7 Conclusion and Future Work

The most important conclusion in this work is that EEG signals outperform non-EEG signals in a person specific setting. In a cross-subject, the performance is similar. This indicates that non-EEG physiological reactions to changes in a person's emotion state are more common between persons. Further research for more advanced transfer learning methods is needed to do cross-subject emotion recognition, as the found performance in this work remained quite low.

REFERENCES

[1] D. O. Bos, "Eeg-based emotion recognition," 2007.

[2] J. Wagner, J. Kim, and e. Andr, "From physiological signals to emotions: implementing and comparing selected methods for feature extraction and classification," *IEEE*, 2005.

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

[4] P. C. Trimmer, E. S. Paul, M. T. Mendl, J. M. McNamara, and A. I. Houston, "On the evolution and optimality of mood states," *behavioral sciences*, vol. 3, pp. 501–521, August 2013.

[5] 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.

[6] unknown, "Electrode placement," 2015.

[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] 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.

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

[11] D. O. Bos, "Eeg-based emotion recognition the influence of visual and auditory stimuli."

[12] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, pp. 2225–2236, 2010.

[13] J. Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," *IEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[14] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter, "Bci competition 2003-data set iib: support vector machines for the p300 speller paradigm," *IEEE Transactions on Biomedical Engineering*, vol. 51, pp. 1073–1076, June 2004.

[15] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for eeg signal classification," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, pp. 141–144, June 2003.

[16] K. P. Bennett and C. Campbell, "Support vector machines: Hype or hallelujah?," *SIGKDD Explor. Newsl.*, vol. 2, pp. 1–13, Dec. 2000.

[17] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.