Integration of Facial Expression Recognition into Emotional Impact Recognition from Video Content

Güven Erkaya

Bachelor's thesis

Integration of Facial Expression Recognition into Emotional Impact Recognition from Video Content

Güven Erkaya

December 4, 2019

Institute for Data Processing
Technische Universität München

Güven Erkaya. *Integration of Facial Expression Recognition into Emotional Impact Recognition from Video Content.* Bachelor's thesis, Technische Universität München, Munich, Germany, 2019.

Supervised by Prof. Dr.-Ing. K. Diepold and Philipp Paukner, M.Sc.; submitted on December 4, 2019 to the Department of Electrical Engineering and Information Technology of the Technische Universität München.

© 2019 Güven Erkaya

Institute for Data Processing, Technische Universität München, 80290 München, Germany, `http://www.ldv.ei.tum.de/`.

This work is licenced under the Creative Commons Attribution 3.0 Germany License. To view a copy of this licence, visit `http://creativecommons.org/licenses/by/3.0/de/` or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Abstract

In this work, video affective content analysis with and without using facial expression recognition is tested as a high-level concept. Regarding this, a publicly available large video dataset is employed too analyze the performance of emotional impact prediction of the videos with the proposed features. For that a facial expression recognition model on images is utilized. As it predicts the facial expressions from images, a method has to be developed to make it work with the videos. Therefore a key frame extraction is carried out. Afterwards, the facial expression are obtained from the detected faces on the key frame images of the videos by the facial expression predictor. For this work, these facial expression are evaluated using machine learning algorithms like linear regression and PLS regression to examine, if the predicted facial expressions can support the emotional impact prediction of the videos as additional features. It is found out that the association between the predictor and predicted values are not appropriate for the used methods. Hence a precise evaluation of the facial expressions from the gathered results wasn't possible. Instead the facial expressions are handled individually for inferring the possible advantages and setbacks of the facial expression prediction in video-induced emotion recognition.

Contents

Abstract **3**

1 Introduction **11**
1.1 Motivation 11
1.2 Research Questions 11
1.3 Scope of the Work 12
1.4 Outline 12

2 Emotions **13**
2.1 The Term "Emotion" 13
2.2 Utilization of Emotion 14

3 A Large Dataset for Predicting the Emotional Impact of the Videos **19**
3.1 Discrete LIRIS-ACCEDE Collection 19
3.2 Annotations 20
3.3 Video Features 21
3.4 Prediction of Affective Scores from Rankings 22

4 Facial Expression Recognition **25**
4.1 Overview 25
4.2 Databases 25
4.3 Route Map 26

5 Experiments **29**
5.1 Aim of the Experiment 29
5.2 Tools 29
5.3 Methodology 32
5.3.1 Retraining of the Facial Expression Predictor 32
5.3.2 Facial Expression Recognition in the Video Key Frames . . . 33
5.3.3 Linear Regression for Emotional Impact Prediction of the Videos 33
5.3.4 PLS Regression for Emotional Impact Prediction of the Videos 34

6 Results and Evaluations **35**
6.1 Evaluation Terms 35
6.2 Results 36
6.2.1 Test on the Facial Expression Predictor 36

6.2.2 Facial Expressions in Key Frame Images 37
6.2.3 Linear Regression 39
6.2.4 PLS Regression 39

7 Conclusion **43**

A Appendix **49**

List of Figures

2.1 Mapping of the categorical emotions in valence arousal space [34]. 18

3.1 Joint quantized histogram of ranks for the 9,800 excerpts in the valence-arousal space. For example, the bottom-left cell shows the number of video clips with a valence and an arousal rank between 0 and 700 [4]. 22

4.1 Visualisation of the 68 facial landmarks of dlib [29]. 27
4.2 Pipeline of training a facial expression recognition model [29]. 27

6.1 Confusion matrix of a test with 10% of the databases for facial expression prediction. 37
6.2 A key frame image with facial landmarks on two founded faces. . . . 38
6.3 A key frame image with facial landmarks. 39
6.4 A key frame image with facial expression angry. 40
6.5 Angry facial expression with facial landmarks. 41
6.6 Arousal predictions vs. ground truth with additional features. 42
6.7 Valence predictions vs. ground truth with additional features. 42

A.1 Arousal predictions vs. ground truth. 49
A.2 Valence predictions vs. ground truth. 50
A.3 Valence correlation map. 51
A.4 Arousal correlation map. 52
A.5 Arousal coefficients for linear regression. 53
A.6 Valence coefficients for linear regression. 54
A.7 Grid Search for PLS regression accounting arousal with emotions. . 55
A.8 Grid Search for PLS regression accounting valence with emotions. . 55

List of Tables

4.1 Used databases to retrain Petrov's model [29]. 26

6.1 Arousal regression results. 40
6.2 Valence regression results. 41

1 Introduction

In this chapter, the research questions that are going to be answered in this work are presented. The motivation for the work and the importance of answering these questions are explained alongside the general scope of the work. The reader can also find an outline of the work.

1.1 Motivation

Emotions are subjective for all human beings. Despite of this subjectivity, it is relatively hard to conceptualize emotion as an effort to utilize it. On the other hand, along with improvements made in computer vision in the last decades, high-level concept detection can be done very efficiently. Thus, applying these high-level concepts in to model emotion accurately became an irresistible field of research in the last years. Trying to tackle the subjectiveness of emotion is the main challenge in this work. If it can be achieved to make a machine learn the emotions of the livings, a major progress in terms of conveniency in our lives can be provieded. In this manner, this work tries to learn and predict the emotions that are induced by videos. Video affective content analysis itself has large number of applications such as mood based video content recommendation, efficient movie browsing and summarization, soundtrack recommendation to make user-made videos more attractive, and optimized insertion of ads in the videos. As working in a field of research where there are no clear guidelines but emotions, determination and durability is needed, to make every failure worthy. Having in mind that if emotions could be modeled accurately, it can amount to a new age of service for humans as then, everything could be conducted according to the best feelings of humans. As an effort to be part of tackling this challenge, I am going to try facial expression recognition in videos to see if an improvement can be provided in predicting the emotional impact of the videos.

1.2 Research Questions

This work seeks for ways for an answer of the following question:

Can the facial expressions detected from videos be used to predict the emotional impact of the videos effectively?

It is to be explored, how one could employ a facial expression recognition model for video-induced emotion prediction. Subsequently it is clarified, if a model working on images can work out well for the videos and how? Most importantly it is analyzed if the facial expression of videos can supply efficiency in the prediction of the impact of emotions in the videos.

1.3 Scope of the Work

As this work consists of emotions and computation of the emotions, it includes a interdisciplinary field of research. It is important to combine the psychological point of views of emotion with the analytical approaches of the computational intelligence effectively. Thereby this work comprises research from video affective content analysis and the biological researches of emotion. Video affective content analysis is itself an interdisciplinary field of research that involves with associating the techniques in cinematography with the techniques in machine learning. Thereby this the intended induced and expected emotions of the cinematography is to be modeled using the features that are carried out from movie and video contents for machine learning practises. This work focuses on predicting the induced emotion of the videos, utilizing the features and methods of these interdisciplinary fields.

1.4 Outline

In chapter 2 the handling of emotions are described. The ways to measure emotion, as well as the ways to portray emotion for a further computation are gone through. Then chapter 3 digs into a large database of videos with emotional annotations and features that are going to be used for testing the facial expression recognition performance. This facial expression recognition model is explored in chapter 4 concisely. Which is followed by experimentation methods and tools in chapter 5. The experimentation results are discussed in chapter 6 along with required evaluation terms. Consequently in the conclusion part the results are examined and interpreted for the reader.

2 Emotions

In this chapter, it is explained how emotions are handled for a computational analysis. An overview is also given from the pschological point of view.

2.1 The Term "Emotion"

> Almost everyone except the psychologists knows what an emotion is [42].

Although the term "emotion" is used very often, the answer following the question "What is an emotion?" is rarely same for different individuals, scientists or laymen alike.

> Fear, terror, anger, rage, horror, lust, embarrassment, agony, excitement, disgust, grief, jealousy, shame, humiliation, amusement, laughter, joy, sorrow, weeping are emotions.

This definition by Young [42] is just one of the hundreds of definitions made by researchers. Kleinginna and Kleinginna state that, a major problem in the field of emotion has been the wide variety of definitions that have been proposed [21].

Many psychologists fail to accept, that there is no commonly, even superficially, acceptable definition of what a psychology of emotion is about [26]. According to English and English [14], emotion is virtually impossible to define except in terms of conflicting theories, though there is a fair agreement in classifying as emotion such phenomena as fear, anger, joy, disgust, pity, affection, etc. Then it becomes imperative to generate a minimal consensus about the defining features of the different types of affective phenomena [36]. In [21], one can find more than hundred definitions of emotion classified into an outline of 11 categories in an attempt to reach a consensus. To be able to guide a research in this field, cumulativeness in the definitions is a must, so that productive comparisons across different results can be made and progress can be achieved. On the other hand, for "computing" emotion effectively, one needs even clearer lines.

For that, the space of emotional states can be objectified by psychologists use of two fundamental methods:(1) the discrete approach and (2) the dimensional approach. It may be seen that it is easy to determine if someone is feeling a certain emotion but as emotional state refers to the dynamic state when one experiences

an emotion [30]; diligence and susceptibility is required to model an emotion in the afore-mentioned methods (1),(2).

To get the discrete and dimensional approach in use, there are different ways in measuring the emotions including self-report measures, observational measures and psychophysiological measures. In the following section and subsections, I am going to give a short overview on these measurement ways to make it clear for the reader how emotion is employed but not go in detail as it's not the topic of this work. One may refer [27], [18], [8] for a deeper uptake. As both of them find a significant use in this work. I am also going to describe the two methods(1),(2) for objectifying emotion.

2.2 Utilization of Emotion

To practice emotion, majorly it is either classified into a certain number of clearly separable states in the discrete method (qualitative) or it's domain is conceptualized by two or three predefined dimensions in the dimensional method (quantitative). Nevertheless, the discrete and dimensional definitions of emotion are related [38]. In fact one can map the categorical emotional states into the dimensional space as seen in figure 2.1 and also in [17]. Before being able to categorize it or determine where the emotion lies in the predefined dimensions, one has to measure it.

Measurement Types

(a) Self-Report

Self-report measures are basicly participants subjective comments on how he/she experiences an emotion. These comments can take forms in online or written questionnaires, phone call or in person interviews and instant elicitations, where participants instantly scale their experienced feelings towards the inspected object by tools like joysticks or rating dials. They are flexible regarding the conduction of the measurements as they can be reported during("online") or after ("offline") perceived stimuli, what makes them very convenient to administer. Besides, they usually require less expensive equipment except for more sophisticated strategies relating with instant elicitations. Regarding these, this type of measurements are the most widely used method to measure emotion [15].In this method both the participants and conductors approaches are crucial, so that the experienced emotions can be reported accurately. Biased responses of participants, memory failures when participants can't accurately remember his/her perceived feelings, poor self perception or participants and researchers disunity on the felt emotion as in participant being unable to report his/her experienced emotion through the used method are the main problems that may occur. According to Robinson and Clore [33], the validity of self reports depends strongly on the administration of the method. To overcome such

problems and the validity issue, different approaches can be taken. E.g. PANAS scales [40] developed by Watson, Clark and Tellegen and PANAS-X [39] developed by Watson and Clark as an expanded version of PANAS-X to increase the validity of self-reports as a solution on the conductors side. Also reliable respondents from the related research field should be preferred for more valid reports. Self-report measures take part in this work as LIRIS-ACCEDE[9] video labels are provided by this method.

(b) Observational Measures

To begin with observational measures, observation is needed first. After observing or recording a persons behaviour for a later observation, it should be possible to infer person's emotional state from vocal characteristics, facial displays, and whole-body behaviours [27]. Whilst observation with or without the help of special inspection instruments, the observed participants behaviour must remain independent from the observation. By this method, participants self awareness problems or biased responses in terms of reporting their experienced emotion can no longer be an issue but the participants experienced emotions may also not correlate accurately with observers indications. Methods as examples for observational measures are Facial Action Coding System (FACS) [13] developed by Ekman and Friesen, which allows users to encode any possible facial muscle movement (observed in photographs, motion picture film or videotype) in terms of anatomically based action units (AU) [12] or Michael's model[29], which detects faces and draws facial landmarks on the faces observed in pictures. Though materials such as audio and video equipment as well as qualitative and observational coding software can be quite costly, as can the required training for some of these methods (e.g., FACS) [18]. One can also use observational methods as a supplement for self reports. Common sense suggests that for obtaining empirical information about emotions, both methods—self report and external observation—are useful [32].Thus issues like biased reports or memory failures can be minimized. For more detail in observational approaches in measuring emotion, one may refer [32] apart from the mentioned [27], [18], [8].

(c) Psychophysiological Measures

Psychophysiology is the branch of psychology that involves with the physiological nature of psychological processes. The most frequently used psychophysiological measures of emotion are indicators of the autonomic nervous system (ANS), including indicators of respiratory functioning (e.g., oxygen uptake), cardiac functioning (e.g., blood pressure, heart rate) and electrodermal measures [18] such as skin conductance level (SCL) or short-duration skin conductance responses (SCRs)[27]. Researchers use such feedbacks to estimate participants affective states. Also some other methods provide emotion-relevant information by assessing central nervous

system (CNS) functioning like electroencephalography (EEG) in which electrical brain activity is measured through noninvasive electrodes placed on the scalp. As this measures are independent of the participants inability to report an experienced emotion, they may overcome the validity issue of self reports. On the other hand, repeated measurements required for more reliable results [18] as well as trained researchers and equipped labs which may be costly in terms of time and money. Interested reader may refer [31] and [2] other than [27], [18], [8] for more detail in administration and utilization of this method.

Measurement Levels: Discrete or Dimensional

(1) Discrete Emotion Approach

Discrete emotion model, also commonly referred to as the categorical emotion model is based on a defined discrete set of core emotions. According to Scherer

> it goes back to the origin of language and the emergence of words and expressions describing clearly separable emotional states"[36]

, which depicts that each of this categorized core emotions is unique in terms of perception, expression and physiology. For more scientific terms, one may refer [10], in which Ekman proposed 11 characteristics that distinguish core emotions or with his words "basic" emotions from each other and from other affective phenomena. According to Ekman [10], the term basic distinguishes those who maintain that there are a number of separate emotions, that differ one from another in important ways. It has also been used to describe elements that combine to form more complex emotions. Moreover Ekman and Friesen [11] provided evidence in their work that, the 6 basic emotions "anger, disgust, fear, sadness, happiness, surprise" are universal, associating with the facial muscular patterns. They showed that the facial behaviours of this emotions can be recognized regardless of ethnic and cultural differences. Many discrete sets of core emotions are proposed but there isn't any accepted final set. In this work Ekman's [10] 6 basic emotions find use, as Michael [29] also utilized these emotions to classify the predicted emotion of faces in his model. In addition to this set of 6 basic emotions, the emotional states "neutral" and "contemptuous" are also counted in Michael's model [29] and therefore used in this work. The categorical approach is applied in many affective movie content analysis work but faces a granularity issue since the number of emotion classes is too small in comparison with the diversity of emotions perceived by film viewers [6]. On the otherhand with increasing emotion classes, fuzziness may occur because of ambiguous issues in personal interpretations and language. As in [4] a dimensional approach may be more proper in this case.

(2) Dimensional Emotion Approach

In dimensional approaches, emotional states are defined according to their position in a continuous n-dimensional space. No consensus has been reached in terms of the labels of dimensions, but firstly introduced by Wundt [41] the three dimensional valence-arousal-tension space (or with his terms "pleasurable or unpleasurable", "arousing or subduing" and "strain or relaxation") is among the most famous ones of this approach. The valence dimension contrasts states of pleasure with states of displeasure ranging from positive valence (e.g., happy, joyous) to negative valence (e.g., sad, disappointed). The arousal dimension contrasts the activation level of the emotion whereas it ranges from low arousal (e.g., passive, calm) to high arousal (e.g., excited, surprised). On the other hand, the tension dimension ranging from tensed to relaxed may differ in researches as some other works like [35] handles dominance as a label instead, typically ranging from no control to full control which identifies if the experienced emotion is either submissive or dominant. As Baveye, Dellandréa, Chamaren and Chen also did in [4], given the difficulty of consistently identifying a third dimension (such as tension, control, or potency) different from arousal, many modern dimensional theorists limit themselves to the valence and arousal dimension [36]. In this case, the emotional feeling of the respondent is described by a point in this valence-arousal space[36]. E.g. the emotional state "happy" lies between positive valence and positive arousal axes while "afraid" lies between positive arousal and negative valence axes as seen in 2.1. Also in this work the valence-arousal dimensions are used, as the Discrete LIRIS-ACCEDE collection[4] is annotated along the valence and arousal axes.

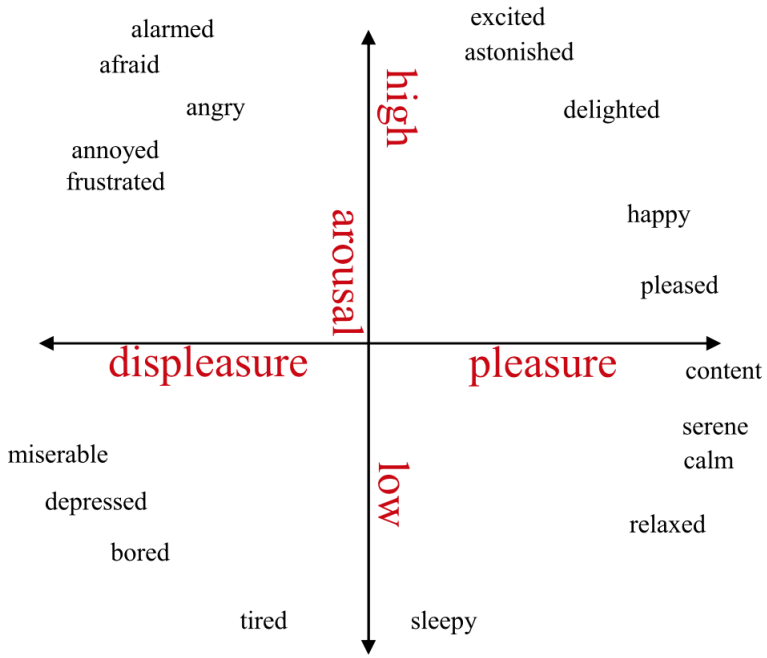

Figure 2.1: Mapping of the categorical emotions in valence arousal space [34].

3 A Large Dataset for Predicting the Emotional Impact of the Videos

Efficient training and benchmarking of computational models, require a large and diverse collection of data annotated with ground truth[1], which is often difficult to collect, particularly in the field of affective computing [9]. The issue in this field is that most of the existing datasets contain less amounts of video resources which are often publicly not available.

As an effort to solve this problem, Dellandréa and his co-researchers created the LIRIS-ACCEDE dataset[9], which is composed of 160 professionally made and amateur movies with different movie genres including horror comedy, drama, action and etc. The movies are shared under Creative Commons[2] licenses and therefore publicly available together with the annotations and the extracted features. The annotations are based on the induced emotions in these movies in terms of valence and arousal. Regarding this, large amounts of data are annotated via crowdsourcing[3] to address the issue of insufficient sized datasets in the field. Dellandréa and his co-researchers also tried to represent a full range of emotions by collecting movies with a large content diversity in LIRIS-ACCEDE [9]. Besides, picking movies without copyright constraints was of their concern.

As almost all investigators dealing with video affective content analysis use their own databases designed for their own goals, LIRIS-ACCEDE was created to be a reliable affective database as an effort to enhance affective computing studies[4]. Moreover, out of the LIRIS-ACCEDE dataset 6 different collections were gathered along with different annotation methods, all addressed to different tasks and experiments. In this work only the Discrete LIRIS-ACCEDE collection [4] is used and described. For information about other collections, interested reader may refer [9].

3.1 Discrete LIRIS-ACCEDE Collection

In Discrete LIRIS-ACCEDE collection Baveye, Dellandrea, Chamaret and Chen [4] extracted 9800 video excerpts from the 160 movies of LIRIS-ACCEDE. While 1000

[1] The term "ground truth" refers to the measurement information of the target variable gained empirically.

[2] https://creativecommons.org/licenses/by-nc-sa/2.0/

[3] Crowdsourcing is using a crowd or a large group of people as a source for obtaining work or practical information via internet.

of the excerpts were segmented manually as a part of testing annotation reliability, others were extracted automatically by robust cut and fade in/out detection algorithms [4]. In a total duration of 26 hours, 57 minutes and 8 seconds, each video clip lasts between 8 and 12 seconds. Furthermore, Baveye and his co-researchers claim that they achieved a great variety of excerpts reflecting the variety of selected movies. For more information about the extraction of the videos one may refer[4].

After the extraction, video clips are annotated in terms of their rankings along the both valence arousal axes. Additionally a large number of features are examined. Thus, two feature sets for both valence and arousal axes are proposed. With these rankings and features, Baveye et al. provided a baseline framework using support vector regression algorithm separately for valence and arousal axes. Different protocols were introduced for testing and evaluating the baseline framework along with the results and all of them are available in[4]. Moreover, the annotated rankings of the video excerpts doesn't represent the distribution of the video clips along the continuous valence arousal dimensions. From the video clip with highest ranked valence or arousal, one cannot understand effectively, how arousing or how joyful this video clip is but can infer that the induced arousal or valence level of the video clip is the highest, relative to other video clips in the dataset. This aspect prevents testing the performance of other models used on other datasets that are annotated in terms of their affective ratings along the valence arousal dimensions. To address this issue Baveye and his co-researchers [3] utilized Gaussian processes for regression to predict the affective scores of each video clip from their rankings.

This regression results are used as valence and arousal values of the video excerpts in this work, along with the extracted features. The annotations, features and the acquirement of the affective scores of the video excerpts are explained in the following sections.

3.2 Annotations

In the annotation process of the discrete LIRIS-ACCEDE video excerpts, more than one million annotations were gathered from more than 1500 trusted annotators using CrowdFlower[4] [4]. Baveye and his co-researchers have decided for a ranking approach in terms of annotation. They claim that asking for pairwise comparisons seems less complex than asking for an absolute value [4]. As the scale of the absolute affective ratings may vary for different annotators, pairwise comparison is chosen instead of ratings to improve the consistency. Regarding this, for each pair of video excerpts, the annotators had to choose which one induced the emotion

[4] CrowdFlower or with it's current name Figure Eight, is an artificial intelligence and machine learning company, which aims to provide high quality data to it's customers via internet, by gathering the demandant and provider together under some conclusive terms . They can be seen as a crowdsourcing company for it's customers. For more information: https://www.figure-eight.com/

most in terms of valence and arousal [4]. As the rankings could also vary in terms of valence and arousal, (i.e. valence and arousal rankings may be different for the same video excerpt) the annotation process was separated. Consequently, annotations along the valence axis and annotations along the arousal axis were collected aside, under same protocols.

Furthermore, to generate the comparisons and rank the video excerpts along the annotations efficiently, quicksort algorithm[5] was used [4]. Nevertheless, in order to make sure that the annotators are doing their job seriously, 100 test question were randomly inserted into the tasks for both axes [4]. This test questions supposed to have obvious answers, therefore the annotators that are failing to answer an amount of test question correctly, were counted as unreliable by CrowdFlower and none of their answers were accepted. Increasing the accuracy of the comparisons was aimed by these test questions.

Regarding these matters, in 3.1 one can see the distribution of the video excerpts in terms of their rankings along valence and arousal axes. Each cell represents the number of the video excerpt with a ranking between the shown values on both axes [4].

3.3 Video Features

After investigating a large number of low-level features, Baveye and his co-researchers divided these features into three modalities: audio, still image and video features [4]. The audio features were extracted from the audio signals of the video excerpts. The still image features on the other hand, were extracted from the key frames of the excerpts, as the key frame corresponds to the frame with the closest RGB histogram(i.e. pixel intensity distribution) to the mean RGB histogram of the whole excerpt using the Manhattan distance[6]. Moreover, the video features represent the information about the composition (number of scene cuts, fades, etc.) and motion [4]. Nonetheless, two different features sets were provided for the valence and arousal values having some of the features in common.

The feature set for valence axis consists of 17 features, while the feature set for arousal axis consists of 12 features. Those are also the features utilized in this work. The valence features are namely: colourfulness, hue count, audio zerocrossing rate, entropy complexity, disparity of most salient points (standard deviation of normalized coordinates), audio asymmetry envelope, number of scene cuts per frame, depth of field, compositional balance, audio flatness, orientation of the most harmonious template, normalized number of white frames, the color energy and color contrast,

[5]Quicksort is an efficient algorithm based on divide and conquer, to put a given array or a list of numbers in an order.

[6]The Manhattan distance between two points is the sum of the absolute differences of their Cartesian coordinates.

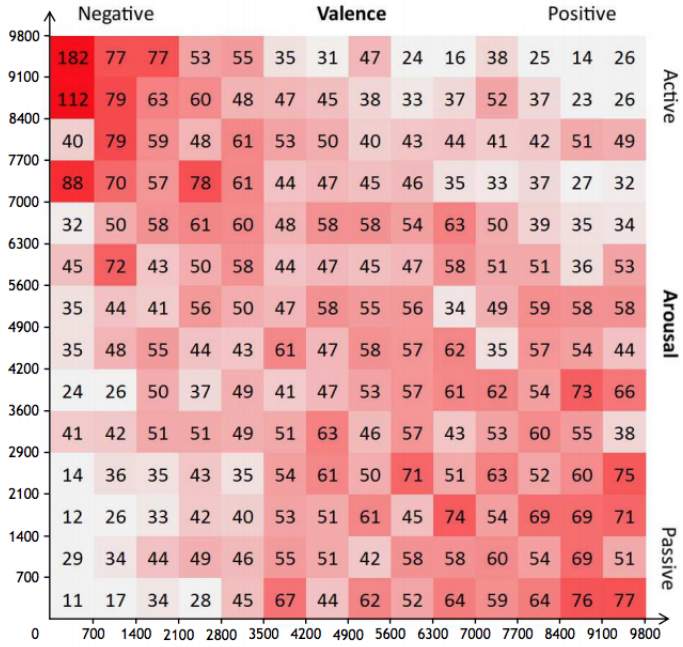

Figure 3.1: Joint quantized histogram of ranks for the 9,800 excerpts in the valence-arousal space. For example, the bottom-left cell shows the number of video clips with a valence and an arousal rank between 0 and 700 [4].

scene complexity (area of the bounding box that encloses the top 96.04 percent of edge energy), number of maximum values in the saliency map and finally, number of fades per frame. Whereas the features for arousal axis are: global activity (average size of motion vectors), standard deviation of the wavelet coefficients of audio signal, the energy corresponding to the most harmonious template, the slope of the power spectrum, median lightness, the lighting feature, length of scene cuts and the audio flatness envelope.

3.4 Prediction of Affective Scores from Rankings

In [5], Baveye and his co-researchers selected 40 video clips of the Discrete LIRIS-ACCEDE collection [4], ensuring that the most reliable film clips in eliciting induced emotions are selected [5]. Afterwards they gathered 28 volunteers to report the induced emotions of this video clips, in terms of scores along the valence and arousal dimensions. This experiment was conducted in a controlled environment and Self-Assessment Manikin [1] was used as a scale for reporting the ratings. It was shown that the controlled affective ratings of this 40 video clips and their va-

lence and arousal rankings are significantly correlated in terms of Spearman's rank correlation[7] thus cross validating the discrete LIRIS-ACCEDE collection [5]. Using this 40 video clips rankings and affective scores, Baveye and his co-researchers [3] trained two Gaussian process regression models for both valence and arousal dimensions, which predict the ratings in a scale between 1 to 5 along the both valence and arousal dimensions of the 9800 video excerpts of Discret LIRIS-ACCEDE collection, from their rankings. In this work, these prediction scores ranging from 1 to 5 are used as labels for both valence and arousal.

[7] Spearman's rank correlation is a measure between two elements in a scale of -1 to 1, that depicts the monotonic relationship of these elements. A correlation close to 1 means as one element increases or the other one also tends to increase or vice versa.

4 Facial Expression Recognition

As emotion is subjective and hard to predict even for us humans, to gain good prediction results in video affective content analysis, low-level features may not be sufficient enough. On the other hand, considering the major progress achieved in computer vision for high-level concepts, it becomes attractive to test them in video affective content analysis with low-level features, aiming to increase the prediction accuracy of the emotions elicited by the videos.

In this chapter Petrov's facial expression recognition model is explored briefly, as it is incorporated with LIRIS-ACCEDE [4] features further on to test facial expression recognition, as a high-level concept in video effective content analysis with low-level features.

4.1 Overview

Petrov [29] made use of 6 databases in his work. The aim of his work was to evaluating the same models performance on different databases. He used support vector classifier as a model and trained and tested it with these databases.

4.2 Databases

In Petrov's work six different databases found use. Five of them consists of portrait images with posed expressions and one includes images of facial expressions in the wild (i.e. non-posed). One of his experiments was for instance to evaluate the performance of model trained on portrait images of posed facial expressions in images with non-posed (i.e. in the wild) facial expressions. It's performance decreases significantly with the non-posed facial expressions in the wild, that's why this database is discarded in this work. Table 4.1 shows the used databases in this work. These databases are publicly available along with the labels of the images in terms of facial expressions seen in 4.1

4.3 Route Map

In his work, [29]Petrov used Histogram of of orientated gradients[1] (HOG) for detecting faces of the labeled images from the databases. After detecting the face, it's bounding box was found and 68 facial landmarks of the face were drawn using dlib library [20] A visualization of this 68 facial landmarks can be seen in figure 4.2. As the facial landmarks vary from image to image, Petrov calculated the euclidean distances of all facial landmarks of a face in an image for more invariant features. Then he proposed the 50 best features in terms of euclidean distances using random forest classifier. Furthermore, he used these 50 best distances as features to feed a support vector machine to predict facial expression in the particular images. An accuracy of 87.94% was achieved after finding the best performing the hyperparameters by a grid search. Reader may see a pipeline of this process in figure 4.2. For this work, a support vector machine is retrained using Michael's hyperparameters in [29]. Further details are in the next chapter.

Table 4.1: Used databases to retrain Petrov's model [29].

Database	Images	Facial Expressions	Subjects
Radboud Faces Database[22]	4824	8 (angry, contemptuous, disgusted, fearful, happy, neutral, sad, surprised)	67
JAFFE[25]	213	7 (angry, disgusted, fearful, happy, neutral, sad, surprised)	10
FEI Face Database[37]	400	2 (happy and neutral)	200
Extended Cohn-Kanade[23]	637	8 (angry, contemptuous, disgusted, fearful, happy, neutral, sad, surprised)	123
KDEF[24]	980	7 (angry, disgusted, fearful, happy, neutral, sad, surprised)	70

[1] Histogram of oriented gradients describe the shape of an object in an image by the distribution of pixel intensity gradients.

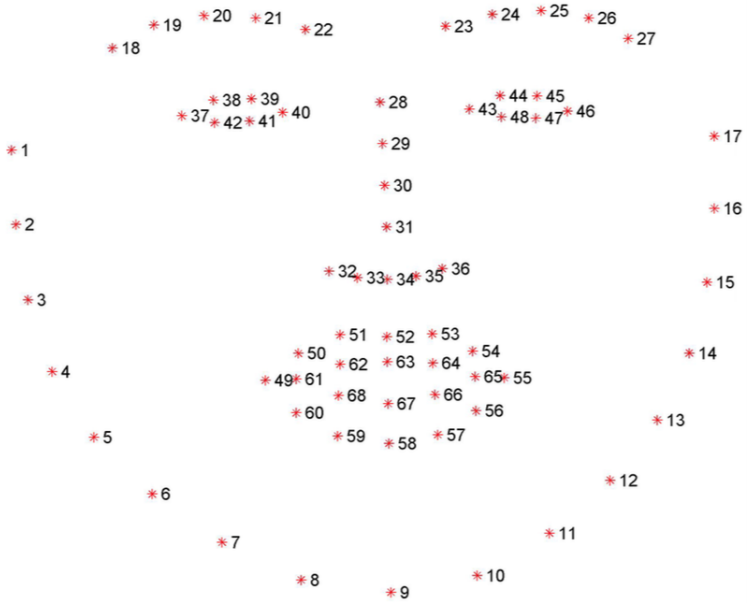

Figure 4.1: Visualisation of the 68 facial landmarks of dlib [29].

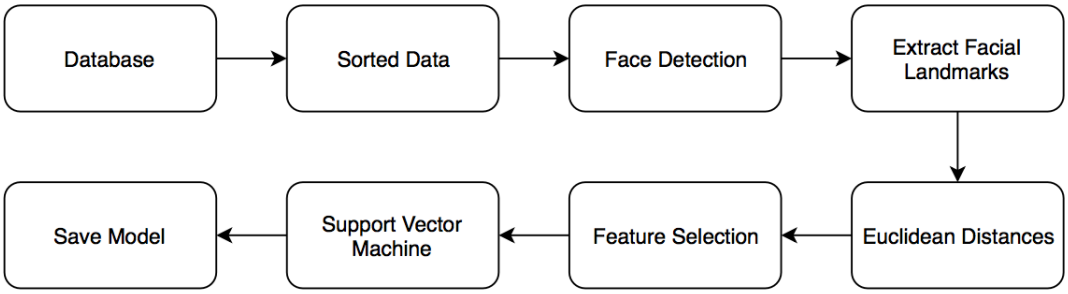

Figure 4.2: Pipeline of training a facial expression recognition model [29].

5 Experiments

In this chapter, the aim of the experiment is depicted firstly. Then the techniques and tools for the experimentation are described. Besides, the followed methodology in the experiments is explored.

5.1 Aim of the Experiment

Before doing the experiments, the goal in this thesis was to test facial expression recognition as an additional feature for predicting the emotional impact of the video excerpts in [9]. To be more specific, it was aimed to analyse, if facial expressions detected in videos are useful features for predicting the induced emotions from the videos. As a part of this analysis, finding the perks and the drawbacks of the facial expressions in predicting the emotional impact of the videos was also an objective. To sum it up, it was to investigate in this work, if the information gathered from the detected faces in the video excerpts could help achieving better accuracies in predicting the emotional impact of the video excerpts.

5.2 Tools

Linear Regression

Linear Regression is a machine learning algorithm commonly used for supervised learning tasks. Its purpose is to predict quantitative outputs in a continuous range. It assumes that the output values and the input values are linearly related to each other, so that a linear approximation between them is reasonable. Thus, it estimates a linear model with the form

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{5.1}$$

of the given input and output pairs, whereas the input vector is represented by $X^\intercal = (X_1, X_2, ..., X_p)$. The goal is to find an approximate model $\hat{f}(x)$ that has the least error representing the relationship between the input/output pairs (x_i, y_i). That is done usually by the least squares method. In the least squares method, it is aimed to minimize the discrepancy between the data pairs and the estimated model(i.e.

residual sum of squares). Regarding this the β in 5.1 is calculated according to the minimization of the residual sum of squares (RSS) as

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - f(x_i))^2 \tag{5.2}$$

$$= \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2. \tag{5.3}$$

After differentiating the RSS with respect to β and setting the first derivative to zero in terms of minimizing it, we get the unique result for 5.1 as

$$\hat{\beta} = (X^{\intercal}TX)^{-1}X^{\intercal}y. \tag{5.4}$$

For more details, interested reader may refer [7], [16].

Partial Least Squares Regression

Partial least squares regression (i.e. PLS regression) is another statistical method in machine learning used for supervised learning tasks. It is different than linear regression models as it seeks for certain directions to decompose the X and Y variables, to obtain features with high variance and also a strong correlation with the responses. These decomposed variables are also called latent variables and after finding the latent variables with the highest covariance, PLS regression applies linear regression on the uncorrelated input directions to predict the output. Meanwhile, the original features are orthogonalized with respect to this directions. Therefore, PLS regression is particularly suitable while working with highly correlated input features as it can reduce the predictors into small uncorrelated set that is highly correlated with responses. Either cross-validation or a grid search is needed to find the directions with the highest variance and correlation. In this work grid search finds use.

Grid Search

Grid search is a machine learning tool for finding the best hyperparameters that are necessary for the particular model. It is not a tool for a model specificly but can be utilized for all machine leanrning models that have characteristic hyperparameters. Unlike the estimated parameters, hyperparamaters are set before the training process begins as their value cannot be estimated from the data. Therefore, a grid search is needed to be done before the training process. In this work, the components that have high variance and high correlation with the response is found by grid

search for the PLS regression. After defining a range of values for the hyperparameters, it iterates through all the defined parameters to find the best performing one. However, it doesn't say anything about how well the model can generalize. For that, a test of the model should be done with the best found parameter on some dataset different from the one used in grid search.

Support Vector Machine (SVM)

Support vector machine is a supervised learning algorithm in machine learning that is used for both classification and regression[28]. In this work it is utilized as a classifier. Regarding this, a support vector classifier defines a separating hyperplane in the feature space, which separates the data into two classes. While doing that, it aims to create the largest margin between the training points for the classes 1 and –1 in the feature space. As the data points of different classes don't always lay perfectly separated from each other in the data space, regularization parameters of the support vector classifier play an important role to have good generalizing fit. In support vector machine the points inside the defined margin(i.e. support points) make fixed contributions to the width of the margin. The parameter C decides if misclassification are going to be allowed to avoid underfitting, as with a increasing number of C the margin between two classes gets smaller and some of the support points may get out of the margin and thus have no influence on the margin which may lead to poor generalization. On the other hand, the parameter γ decides, how far a data point other than support points in the margin can have an influence on the classification problem. Gamma helps the model to learn the shape of the data, so that the model can generalize accurately while predicting some new data.

Also, along with the regularization parameters, a kernel function needs to be selected for the maximization problem of the margin between to classes. The kernel function maps the the raw data into a high dimensional separable feature space by calculating the inner products of the data. In this work radial basis function is utilized as a kernel for the support vector machine that predicts the facial expression of the detected faces in images. It has the form

$$K(x_i, x_j) = exp(-\gamma\|(x_i - x_j)^2\|), \gamma > 0. \tag{5.5}$$

As in the facial expression recognition model there are 8 different classes for the features, a multi-class approach was needed. Therefore, the "one-against-one" approach in [28] was used, in which for n number of classes $nclass * (nclass - 1)/2$ classifiers are constructed and each of the classifiers are trained with 2 classes. For more details, interested reader may refer [7], [16].

LIRIS-ACCEDE

As my goal is to test facial expression recognition in emotional impact prediction of the videos, a baseline model is needed to compare results with and without facial expression recognition. Thanks to Baveye et al. [4], 9800 videos along with their annotations and features were publicly shared as mentioned in chapter 3, so that I could get them in use to work out a baseline model.

Facial Expression Recognition Model

To test facial expression recognition as a high-level feature in emotion prediction, Michael's model [29] described in chapter 4 is utilized to detect faces and to acquire the facial expressions of the detected faces in the annotated video excerpts.

FFMPEG

To apply the facial expression recognition in the video excerpts, feeding the facial expression recognition model with the key frames of the video excerpts is the idea in this work. For that, a key frame extraction of the video excerpts is needed. Regarding this, FFMPEG[1] is a convenient tool for working with videos. It is a multimedia framework to operate on video, audio, and other multimedia files and streams. In this work it finds use, as some video filters(e.g. scene changes) can be applied to the video excerpts using FFMPEG, to extract the representative key frame images of the video excerpts.

5.3 Methodology

5.3.1 Retraining of the Facial Expression Predictor

To get the facial expression predictor in use for the key frames of the video excerpts, it has to be trained diligently, as so the facial expressions of the detected faces in the key frames could be predicted accurately by it for the tests. Thus a support vector classifier is retrained by combining all of the databases that can be seen in the table 4.1. Petrov's grid search results, as well as his proposal of 50 best distances [29] are taken into account as hyperparameters and features to retrain the support vector machine. After the training process, the support vector classifier is set to predict one of the eight facial expressions angry, contemptuous, disgusted, fearful, happy, neutral, sad or surprised for a face detected on a particular image, if the 50 best distances of the faces facial landmarks are given correctly to the support vector classifier as input variables.

[1] https://www.ffmpeg.org/

5.3.2 Facial Expression Recognition in the Video Key Frames

After retraining the facial expression recognition model, it is ready to be used for facial expression prediction from video key frames. For that, video key frames are extracted using FFMPEG. The extraction resulted in 31261 key frame images out of 9800 video clips. Afterwards, face detection is applied on all of the extracted video key frames to find the 68 facial landmarks $(I_1^\intercal, I_2^\intercal, I_3^\intercal ..., I_{68}^\intercal)^\intercal$ for each of the faces found on the key frames. Whereas, each facial landmark contains it's x- and y-coordinates of its spot in the picture,

$$I_i = (x_i, y_i)^\intercal, \qquad x_i, y_i \in \mathbb{N}_0$$

and the euclidean distances are calculated according to these coordinates as

$$d_{ij} = d(I_i, I_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad \forall i, j \in [1, 68] \;\; and \;\; i \neq j.$$

Then the 50 best performing euclidean distances in Petrov's model [29] are chosen to feed the trained support vector machine to get the facial expression labels as predictions of the trained model. Moreover, to utilize the facial expressions of the detected faces as features, the probabilities of the possible outputs of founded faces are calculated based on the 8 emotions. If there isn't any face found in any of the key frames of a video excerpt, a zero vector is passed instead of the probabilities as continuous facial expression features for that video excerpt. If more than one faces are founded for a video clip, the means of the faces probabilities for each emotion is calculated, so that there is only one feature vector for each video clip in terms of facial expression probabilities. The probabilities range from 0 to 1 showing how certain the support vector classifier is while predicting the facial expression of the faces in the key frames.

5.3.3 Linear Regression for Emotional Impact Prediction of the Videos

To test the video-induced emotion prediction performance with and without facial expression features, linear regression is applied only on the provided LIRIS-ACCEDE [9] features first, to predict the affective scores described in section 3.4. As the scores for valence and arousal are depending on different features, two different linear regression models are trained. This models are then used as baseline models for comparing the prediction results of the models that are trained and tested with additional facial expression features. To evaluate the results for each model, the datasets are manually divided into two for the training set and test set and it is made sure that both the training and the test set include 4900 excerpts of the 80 movies in all models. To sum it up, 4 different linear regression models are trained and tested: two for predicting valence with and without facial expression features and two for predicting arousal with and without facial expression features.

5.3.4 PLS Regression for Emotional Impact Prediction of the Videos

As with PLS regression a more robust handle of the input variables can be achieved, PLS regression models are also trained along with linear regression, to avoid the setbacks that may occur because of the correlations between the high number of features. For the training of the PLS regression model, hyperparameter searches are performed using grid search. Regarding this, four grid searches are conducted for four different models that are tested and evaluated identically as the linear regression models. In this sense, best performing parameters are found by fitting the half of the data(i.e. training set) into the models and the models with best performing parameters are then tested on the other half of the data(i.e. test set). Again in this experiment, two models for predicting valence and two models for predicting arousal are implemented for a clear comparison just like the linear regression models.

6 Results and Evaluations

In this chapter the results of the conducted experiments are evaluated and discussed. Also the evaluation terms are explained concisely.

6.1 Evaluation Terms

Confusion Matrix

Confusion matrix is a table to measure how well a classification model performs. The rows of the matrix represent the true labels of the model, whereas the columns are the predicted values. Obtaining higher frequencies in the main diagonal of the matrix and low frequencies for all other elements of the matrix would show that a good classification is achieved as the diagonals represent how much of the predicted values are matching with the true labels. On the other hand, frequencies in the other elements of the matrix reveal how much of the predictions are misclassified.

R-squared Value

R-squared value or coefficient of determination, is a measure that represents how close the predicted values are to the observed values. It ranges from -1 to 1 and a R^2 score of 1 indicates that the regression model predicts the dependent variables perfectly. It is defined as the proportion of the variance in the target variables that can be predicted accurately by the input variables. For the mean value of $y = (y_1, ..., y_n)^\intercal$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

the total sum of squares is defined as

$$SS_{tot} = \sum_{i=1}(y_i - \bar{y})^2$$

and with the predicted values $f_i = (f_1, ..., f_n)^\intercal$ the sum of squares of residuals is defined as

$$SS_{res} = \sum_{i=1}(y_i - f_i)^2,$$

so that the coefficient of determination is

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}.$$

Mean Squared Error

Mean squared error is a risk metric that defines the expected value of squared error[28]. In this work the error corresponds to the difference between the predicted valuaes $f_i = (f_1, ..., f_n)^\intercal$ on the regression line and the observed values $y = (y_1, ..., y_n)^\intercal$ and it is calculated as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2.$$

Since the mean squared error reduces with a model closer to the observed target values, a smaller value of the mean squared error is aimed for a good performing model.

Pearson Correlation Coefficient

Pearson correlation coefficient or as commonly referred Pearson's r measures the strength of linear relationship between two variables. It ranges from -1 to 1, whereas 1 represents the perfect linear relationship and -1 represents the perfect negative linear relationship between two variables. A Pearson correlation coefficient of 0 means that there is no linear relationship between two variables. It is calculated as follows;

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

with the sample means $\bar{x}$ and $\bar{y}$ of the given data pairs $[(x_1, y_1), ..., (x_n, y_n)]$.

6.2 Results

6.2.1 Test on the Facial Expression Predictor

To see the performance of the facial expression predictor, all the databases in table 4.1 are combined and splitted into a training and test set first. Thus, 10% of data is used as test set for evaluating the facial expression predictor. As it can be seen in figure 6.1, the test performance of the facial expression predictor with the hyperparameters proposed by Petrov [29] is encouraging to apply it on the key frame images. The high accuracies in the main diagonal show that almost all of the facial expressions in the test set are predicted correctly. After visualizing the performance,

the support vector machine is trained again without splitting the dataset into test and training set, so that it can learn from more data to generalize well on the key frame images.

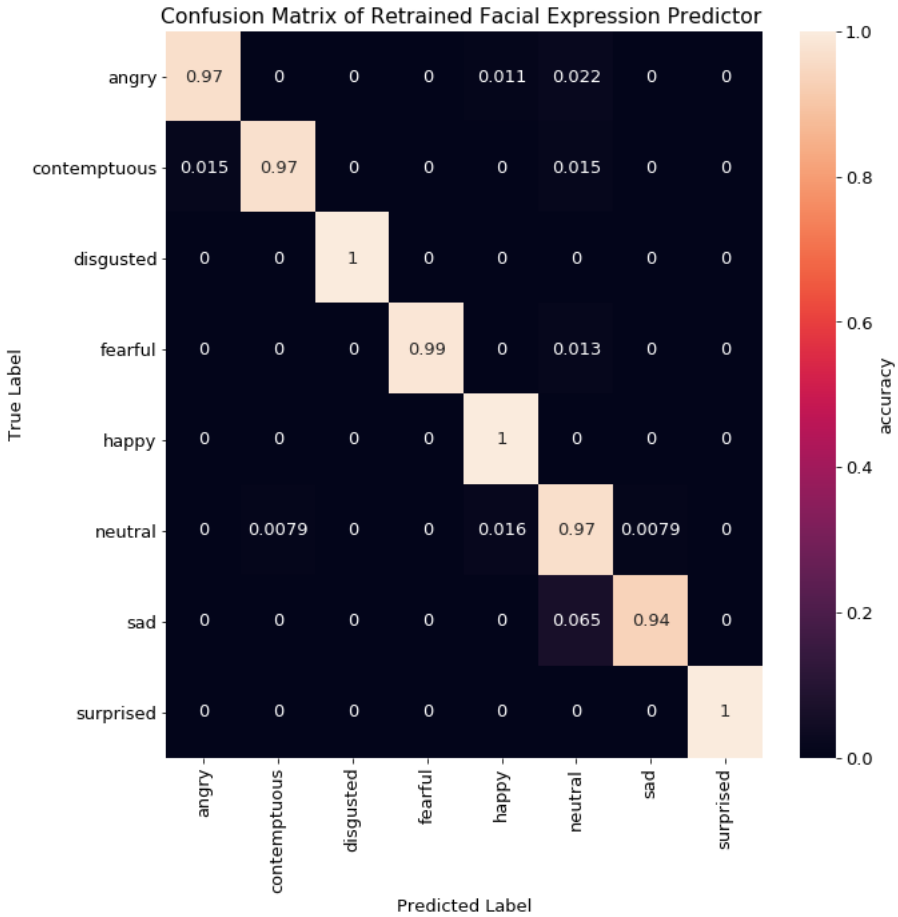

Figure 6.1: Confusion matrix of a test with 10% of the databases for facial expression prediction.

6.2.2 Facial Expressions in Key Frame Images

After applying facial expression recognition on the key frames, the facial expressions of the faces in the key frame images could be predicted. Related to this, 10622 faces are found in the 31261 extracted key frame images. Figure 6.2 shows a key frame of a video clip, which has an arousal value of 4.47 and a valence value of 1.53. The first face detected on the key frame has the facial expression angry, while the second face is reported as sad by the facial expression predictor. Considering the relation-

ship between the discrete emotion approach and dimensional emotions approach, the facial expressions angry and sad can actually correspond the mentioned valence and arousal values as a valence value of 1.53 represents a negative valence value in a scale from 1 to 5 while an arousal value of 4.47 represents a highly activated emotional state. However, to gather effective results in predicting the emotional impact of the videos, it must be considered, what these facial expressions mean for the viewers.

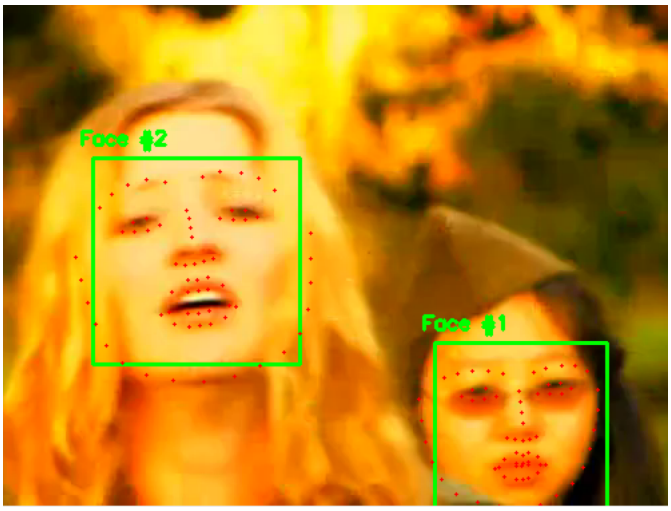

Figure 6.2: A key frame image with facial landmarks on two founded faces.

By minding the emotion probabilities instead of just one emotion per image as features, more features can be gathered based on emotion probabilities, if it is necessary. When one looks at the key frame in figure 6.3, it might be hard to specify a facial expression for the face detected on the key frame. In this work, the facial expression predictor outputs 2 facial expression features in this case. According to the model, the face is angry but has a 0.2 probability of being sad. Thus, this work takes also the feature sadness into account for predicting the affective valence and arousal scores. Though it can still remain unclear, what kind of an impact this face has on the viewer in terms of induced emotion to the viewer.

Considering the emotion induced by a face in the video excerpts, not all of the facial expression contribute to predict the emotional impact of the videos. Figure A.8 shows a key frame of a video clip with a detected face on it. The facial expression predictor outputs angry after detecting this face. Nevertheless, this video clip has a valence score of 3.27 and an arousal score 1.34. It includes an ordinary conversation of two people in a calm atmosphere, which suits the affective scores of negative arousal and neutral valence. As the facial expression of a performer may vary in each frame of a video, it is important that the key frame and the facial expression of the face detected on that particular key frame are matched in a way that the facial

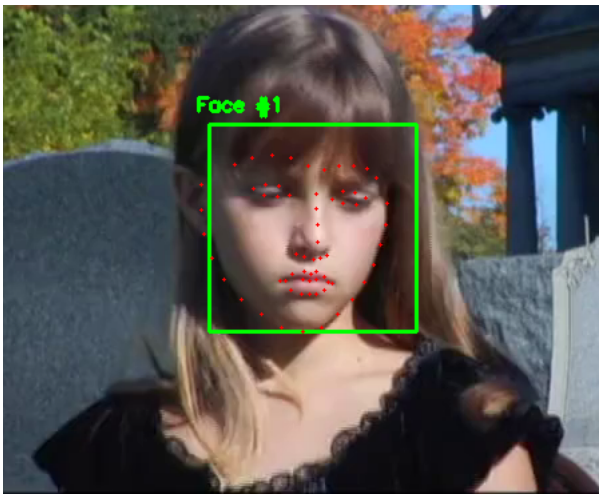

Figure 6.3: A key frame image with facial landmarks.

expression has an impact on induced emotion of the video, so that it can be used as a feature to predict the induced emotions of the videos effectively.

6.2.3 Linear Regression

The results of the linear regression models with and without facial expression for both valence arousal dimensions show that the features provided in LIRIS-ACCEDE [9] and also the additional features extracted from the key frames of the videos in terms of the facial expressions are less likely to be linearly related with the target values of valence and arousal. The difference between the predicted arousal values and the real arousal values can be seen in figure 6.6. The distribution of the real arousal values on the marginal histogram plot shows how unequally the target variables are distributed, which may cause a bad generalizing model. Also with valence it is not different as it can be seen in figure 6.7. The performances of the linear regression models without additional features aren't better than the results with tha additional facial expression features and therefore the resulting plots are included to the appe mentioned here. The coefficient of determination, mean squared error values and Pearson's correlation coefficients can be seen in tables 6.1 and 6.2 for both valence and arousal.

6.2.4 PLS Regression

As the target values depend on a large number of independent variables, PLS regression is used in this work to reduce the correlation of the input features whilst increasing the correlation between target values and the input features. Unfortu-

Figure 6.4: A key frame image with facial expression angry.

nately, no better results are obtained than linear regression for all models in valence and arousal, which can be seen in 6.1 and 6.2

Table 6.1: Arousal regression results.

	R^2	Mean Squared Error	r_p
Linear Regression without Emotions	0.08	0.83	0.29
Linear Regression with Emotions	0.08	0.83	0.29
PLS Regression without Emotions	0.08	0.83	0.29
PLS Regression with Emotions	0.08	0.83	0.29

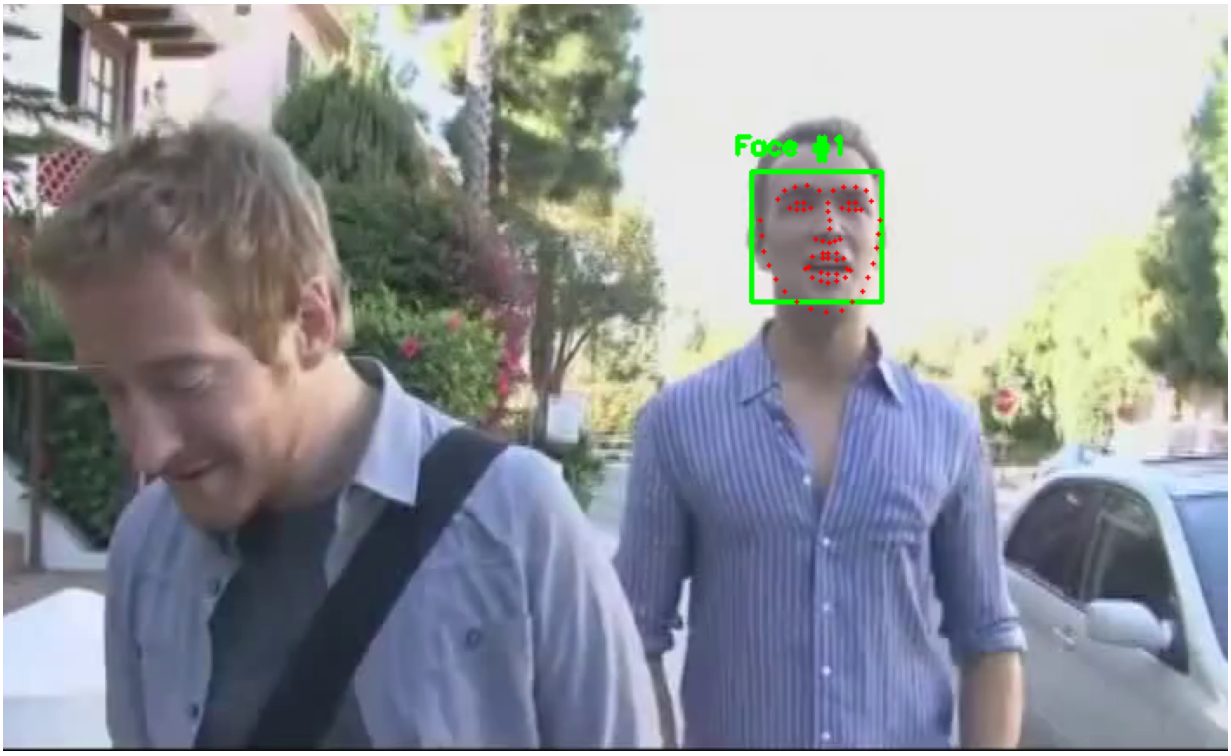

Figure 6.5: Angry facial expression with facial landmarks.

Table 6.2: Valence regression results.

	R^2	Mean Squared Error	r_p
Linear Regression without Emotions	0.07	0.37	0.27
Linear Regression with Emotions	0.08	0.37	0.28
PLS Regression without Emotions	0.07	0.37	0.27
PLS Regression with Emotions	0.08	0.37	0.28

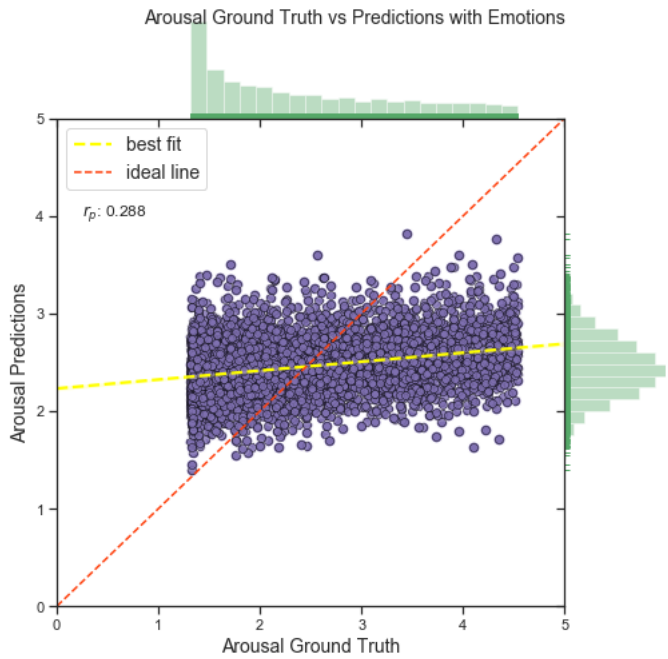

Figure 6.6: Arousal predictions vs. ground truth with additional features.

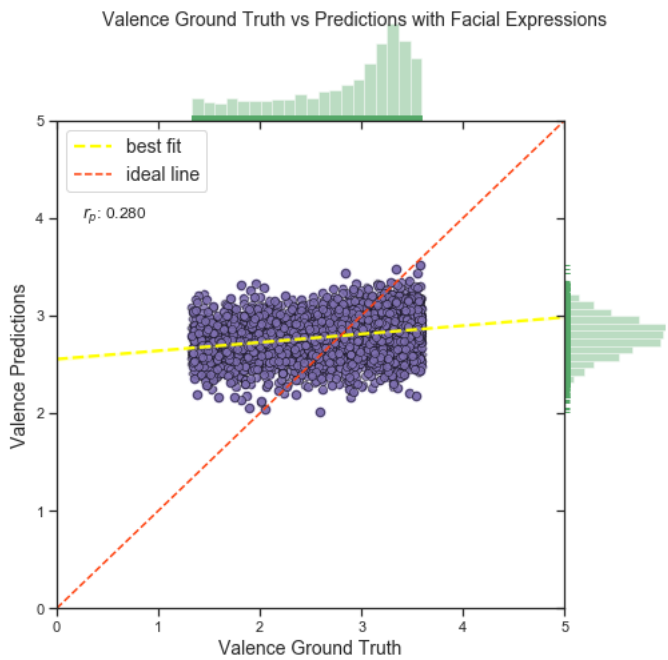

Figure 6.7: Valence predictions vs. ground truth with additional features.

7 Conclusion

Different regression models are applied separately on definite numbers of features to predict the induced valence and arousal from 9800 video excerpts extracted from 160 movies. It is found that both of linear regression and PLS regression don't result well for prediction of valence and arousal scores by the provided features in LIRIS-ACCEDE [9]. Also, it is visualized that the linear correlation between the target values and independent variable are quite low, which depicts that the linear regression method may not be appropriate for emotional impact prediction from LIRIS-ACCEDE videos [9]. On the other hand, the video excerpts are annotated according to their rankings between them and the feature selection is made by selecting the best performing features for prediction of the rankings. So one can't be sure if the best performing features for predicting the rankings of the video excerpts in LIRIS-ACCEDE [4] are beneficial features for predicting the affective scores in terms of valence and arousal. Besides, it is important that whilst the annotation process of the video excerpts, the rankings are done according to the induced emotion from the videos but not the intended emotion in the videos. If failures occur during the annotation process, predicting the emotional impact of the videos wouldn't make sense, as no suitable target values could be gathered. Furthermore, emotional impact predictions couldn't be achieved with high accuracy, so a precise evaluation of the facial expression recognition model for predicting the the emotional impact of the videos would be misleading. In principle, facial expressions recognized in the videos may not be the same as the induced emotion in these particular scenes. A laughing villain may induce the emotions disgust and anger for viewer but a facial expression recognition model would predict happiness as a facial expression. Also not all the facial expressions predicted by a facial expression recognition model have an impact on the induced emotion of the video. Facial expression recognition model can be tried as a helping tool to predict emotional impact of the videos. Combining the facial expression recognition with a emotion recognition model from speech is an idea for further research. Utilizing the features of the speeches can solve some problems in facial expression recognition, as the model then would take also the words of a smiling villain which then can be interpreted as an emotional impact of fear or sadness. In general emotions have high variations in terms of both expressing and receiving. But as the high-level concept detection can be done very efficiently, using the concepts to tackle the emotion recognition problem is inevitable.

Bibliography

1. Measuring emotion: The self-assessment manikin and the semantic differential. In *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), pp. 49–59, 1994. doi:https://doi.org/10.1016/0005-7916(94)90063-9.

2. M.Z. Baig and M. Kavakli. A survey on psycho-physiological analysis measurement methods in multimodal systems. In *Multimodal Technologies and Interact*, 3(37), p. 20, May 2019. doi:10.3390/mti3020037.

3. Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. From crowdsourced rankings to affective ratings. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6. July 2014. doi:10.1109/ICMEW.2014.6890568.

4. Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen. Liris-accede: A video database for affective content analysis. In *IEEE Transactions on Affective Computing*, 6(1), pp. 43–55, Jan 2015. doi:10.1109/TAFFC.2015.2396531.

5. Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen. A Protocol for Cross-Validating Large Crowdsourced Data: The Case of the LIRIS-ACCEDE Affective Video Dataset. In *CrowdMM '14 Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pp. 3–8. Orlando, United States, November 2014. doi:10.1145/2660114.2660115. URL `https://hal.archives-ouvertes.fr/hal-01313188`.

6. Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen. Affective video content analysis: A multidisciplinary insight. In *IEEE Transactions on Affective Computing*, 14(8), 2017. doi:10.1109/TAFFC.2017.2661284.

7. C.M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

8. D. Ciuk, A. Troy, and M. Jones. Measuring emotion: Self-reports vs. physiological indicators. In *SSRN Electronic Journal*, January 2015. doi:10.2139/ssrn.2595359.

9. E. Dellandréa, M. Huigsloot, L. Chen, Y. Baveye, Z. Xiao, and M. Sjöberg. Predicting the emotional impact of movies. URL `https://sigmm.hosting.acm.org/2018/12/18/predicting-the-emotional-impact-of-movies/#6`.

10. P. Ekman. Basic emotions. In T. Dalgleish and M.J. Powers (eds.), *Handbook of Cognition and Emotion*, pp. 45–60. John Wiley and Sons Ltd., 1999.

11. P. Ekman and W.V. Friesen. Constants across cultures in the face and emotion. In *Journal of Personality and Social Psychology*, 17(2), pp. 124–129, 1971.

12. P. Ekman and W.V. Friesen. Measuring facial movement. In *Environmental Psychology and Nonverbal Behavior*, 1(1), pp. 56–75, 1976. doi:https://doi.org/10.1007/BF01115465.

13. P. Ekman and W.V. Friesen. *Facial Action Coding System: Manual*. Palo Alto, Consulting Psychologists Press, 1978.

14. H.B. English and A.C. English. *A Comprehensive Dictionary of Psychological and Psychoanalytical terms; A GUIDE TO USAGE*. New York Longmans, Green, 1973.

15. J. Harley. Measuring emotions. In I.S. Tettegah and M. Gartmeier (eds.), *Emotions, Technology, Design, and Learning*, pp. 89–114. London, UK: Academic Press, Elsevier, 2015.

16. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

17. H. Hoffmann, A. Scheck, T. Schuster, S. Walter, K. Limbrecht, H. Traue, and H. Kessler. Mapping discrete emotions into the dimensional space: An empirical approach. pp. 3316–3320. October 2012. doi:10.1109/ICSMC.2012.6378303.

18. S. Kaplan, R. Dalal, and J. Luchman. Measurement of emotions. In R.R. Sinclair, M. Wang, and L.E. Tetrick (eds.), *Research Methods in Occupational Health Psychology: Measurement, design, and data analysis*, pp. 61–75. Routledge/Taylor Francis Group, 2012.

19. S. Keshav. How to read a paper. In *SIGCOMM Comput. Commun. Rev.*, 37(3), pp. 83–84, July 2007. ISSN 0146-4833. doi:10.1145/1273445.1273458. URL `http://doi.acm.org/10.1145/1273445.1273458`.

20. D.E. King. Dlib-ml: A machine learning toolkit. In *J. Mach. Learn. Res.*, 10, pp. 1755–1758, December 2009.

21. P.R.J. Kleinginna and A.M. Kleinginna. A categorized list of emotion definitions with suggestions for a consensual definition. In *Motivation and Emotion*, 5(4), pp. 345–379, December 1981. doi:10.1007/bf00992553.

22. O. Langner, R. Dotsch, G. Bijlstra, D.H.J. Wigboldus, S.T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database.

In *Cognition and Emotion*, 24(8), pp. 1377–1388, 2010. doi:10.1080/02699930903485076.

23. P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. pp. 94 – 101. July 2010. doi:10.1109/CVPRW.2010.5543262.

24. D. Lundqvist, A. Flykt, and A. Öhmann. The karolinska directed emotional faces (kdef)-cd rom from dept. 1998.

25. M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 200–205. April 1998. doi:10.1109/AFGR.1998.670949.

26. G. Mandler. Emotion. In E. Hearst (ed.), *The First Century of Experimental Psychology*, pp. 275–321. Lawrence Erlbaum Associates, Hillsdale New Jersey, 1979.

27. I.B. Mauss and M.D. Robinson. Measures of emotion: A review. In *Cognition emotion*, 23(2), pp. 209–237, 2009. doi:10.1080/02699930802204677.

28. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python . In *Journal of Machine Learning Research*, 12, pp. 2825–2830, 2011.

29. M. Petrov. *Implementation and Evaluation of a Facial Expression Recognition System*. Technical Report, Technical University Munich Chair for Data Processing, Arcissstrasse 21, 80333, Munich, Germany, 2019.

30. R.W. Picard. *Affective Computing*. Technical Report, M.I.T Media Laboratory Perceptual Computing Section, 20 Ames St., Cambridge, MA, USA, 1997.

31. R.W. Picard, E. Vyzas, and J. Healey. Toward machine emotional intelligence: Analysis of affective physiological state. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(10), pp. 1175–1191, October 2001. doi:10.1109/34.954607.

32. R. Reisenzein, M. Junge, M. Studtmann, and O. Huber. *Observational approaches to the measurement of emotions*, pp. 580–606. April 2014.

33. M.D. Robinson and G.L. Clore. Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. In *Journal of Personality and Social Psychology*, 83(1), pp. 198–215, 2002. doi:10.1037//0022-3514.83.1.198.

34. J. Russell. A circumplex model of affect. In *Journal of Personality and Social Psychology*, 39(6), pp. 1161–1178, 12 1980. doi:10.1037/h0077714.

35. J. Russell and A. Mehrabian. Evidence for a three-factor theory of emotions. In *Journal of Research in Personality*, 11(3), pp. 273–294, September 1977. doi:10.1016/0092-6566(77)90037-X.

36. K.R. Scherer. What are emotions? and how can they be measured? In *Social Science Information*, 44(4), pp. 695–729, December 2005. doi:10.1177/0539018405058216.

37. C.E. Thomaz and G.A. Giraldi. A new ranking method for principal components analysis and its application to face image analysis. In *Image and Vision Computing*, 28(6), pp. 902–913, 2010. doi:https://doi.org/10.1016/j.imavis.2009.11.005.

38. S. Wang and Q. Ji. Video affective content analysis: A survey of state-of-the-art methods. In *IEEE Transactions on Affective Computing*, 6(4), pp. 410–430, October 2015. doi:10.1109/TAFFC.2015.2432791.

39. D. Watson and L. Clark. The panas-x: Manual for the positive and negative affect schedule-expanded form. In *Psychology Publications*, January 1999. doi:10.17077/48vt-m4t2. URL `https://ir.uiowa.edu/psychology_pubs/11`.

40. D.A. Watson, L.A. Clark, and A. Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. In *Journal of personality and social psychology*, 54(6), pp. 1063–1070, 1988.

41. W.M. Wundt. *Outlines of psychology*. Leipzig, W. Engelmann; New York, G.E. Stechert, 1896.

42. P.T. Young. Feeling and emotion. In B.B. Wolman (ed.), *Handbook of General Psychology*, p. 749. Englewood Cliffs, N.J., Prentice-Hall, 1973.

A Appendix

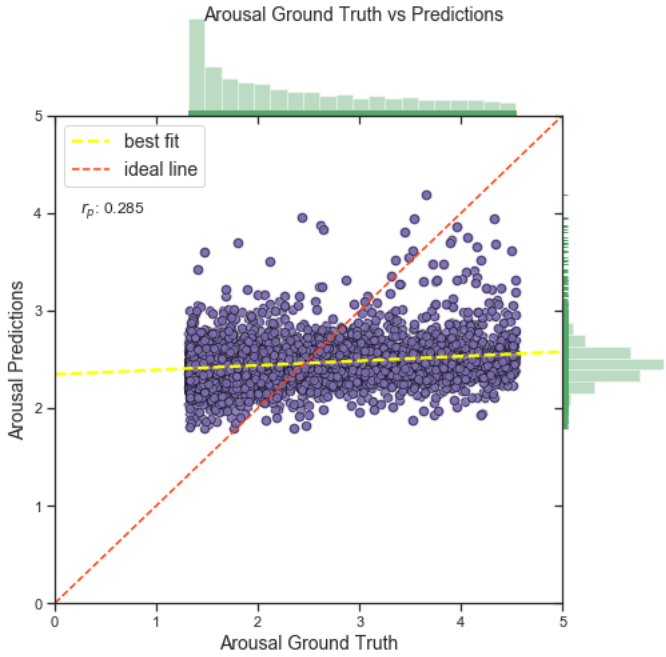

Figure A.1: Arousal predictions vs. ground truth.

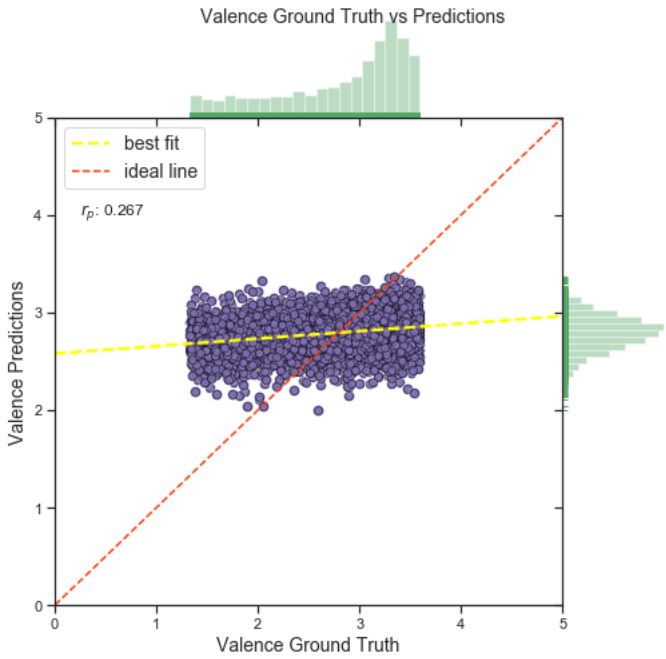

Figure A.2: Valence predictions vs. ground truth.

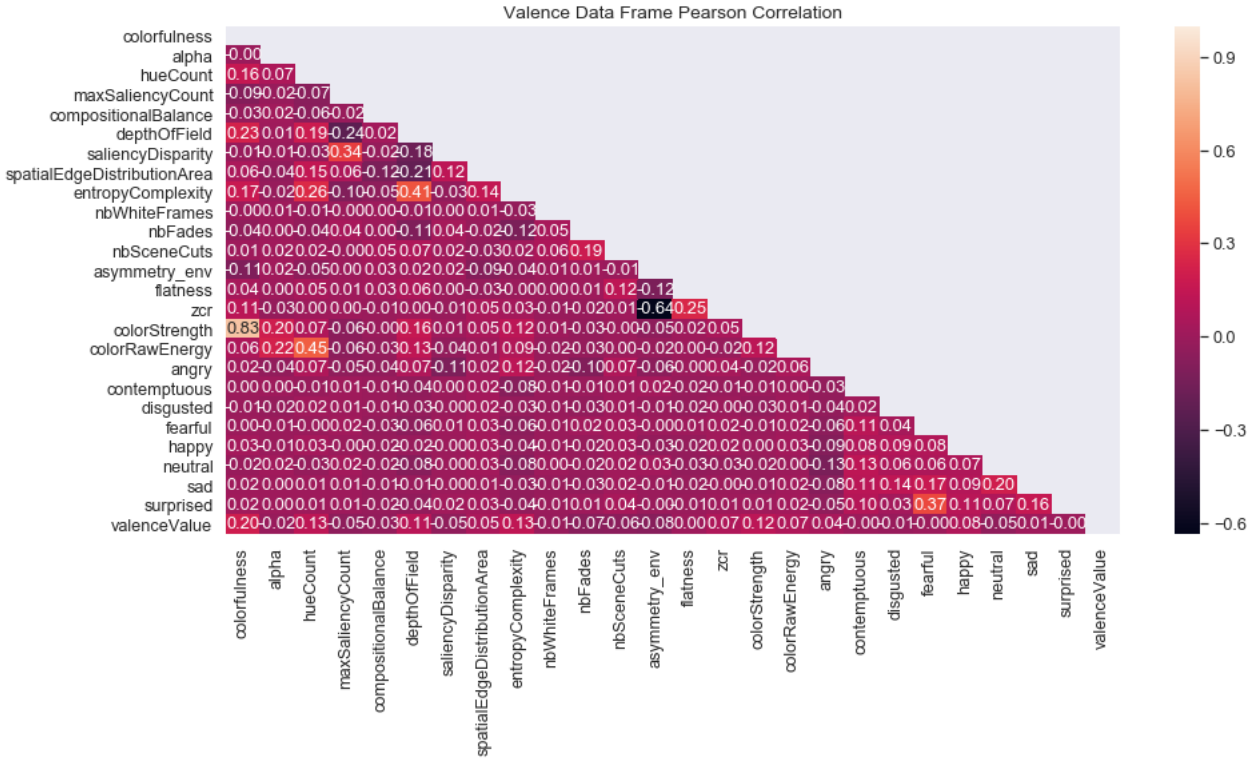

Figure A.3: Valence correlation map.

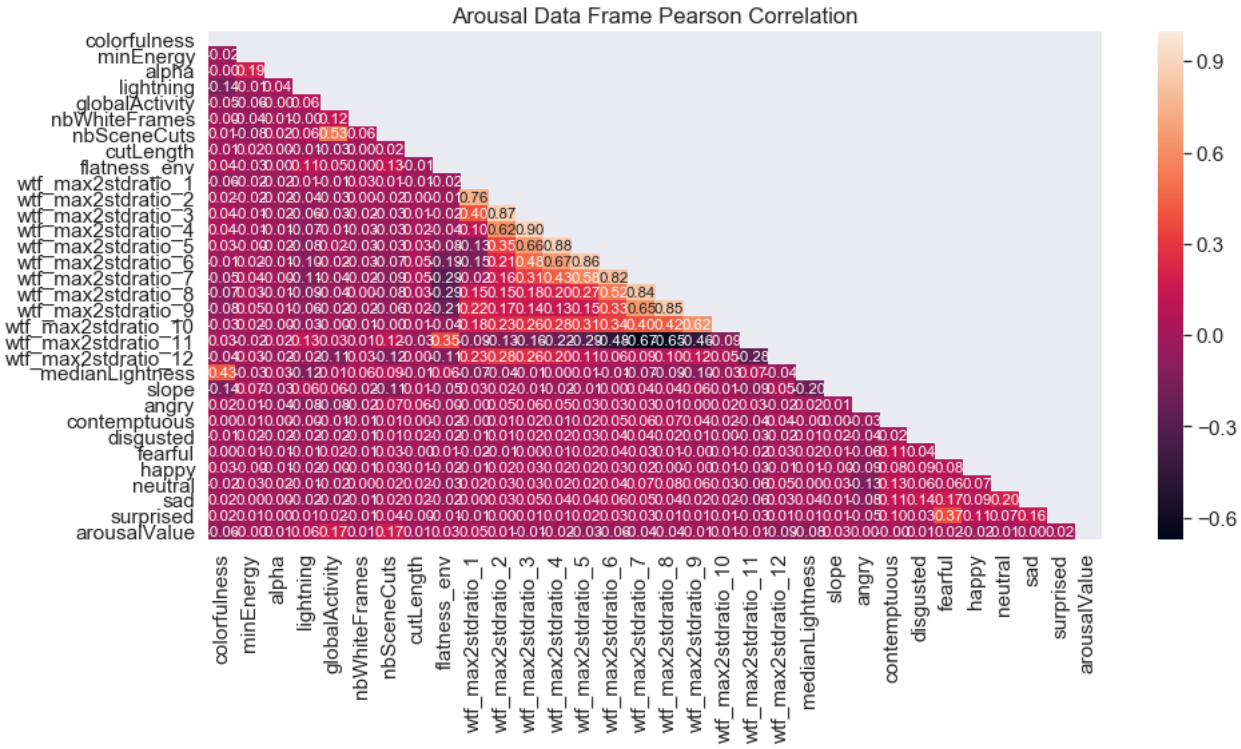

Figure A.4: Arousal correlation map.

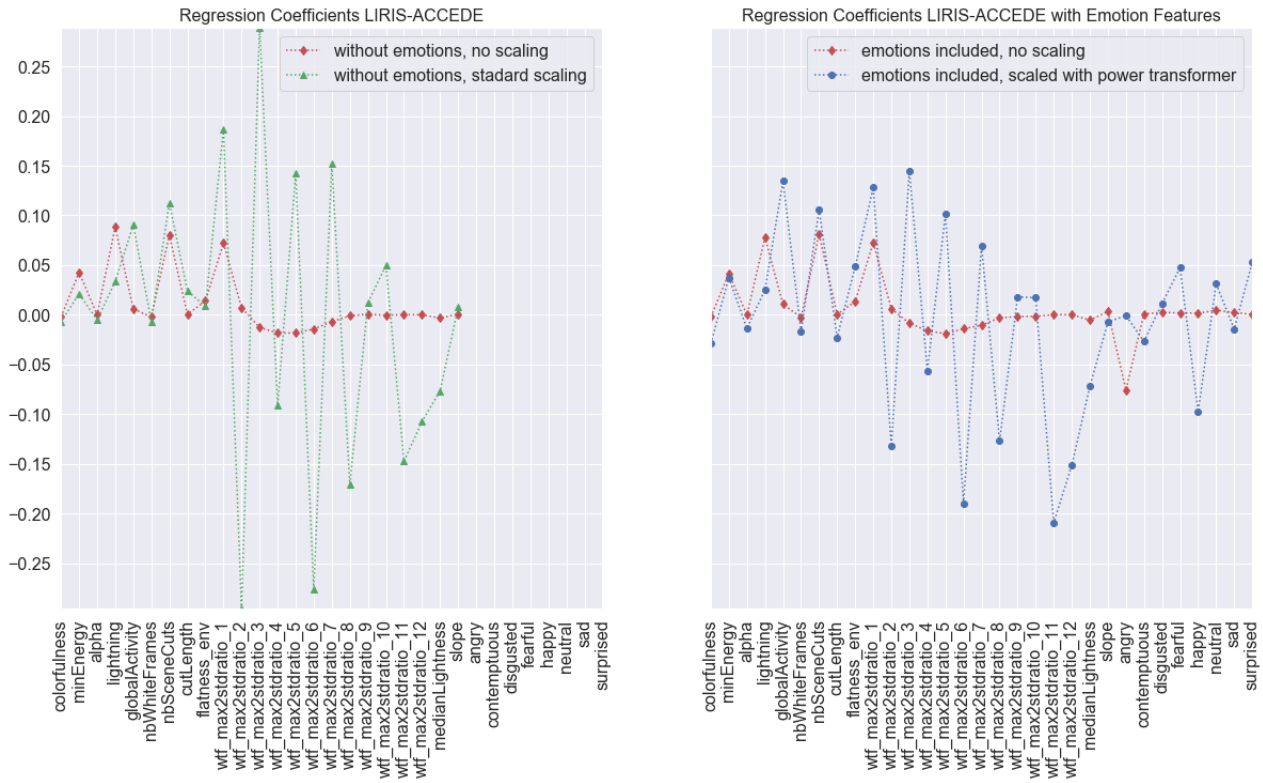

Figure A.5: Arousal coefficients for linear regression.

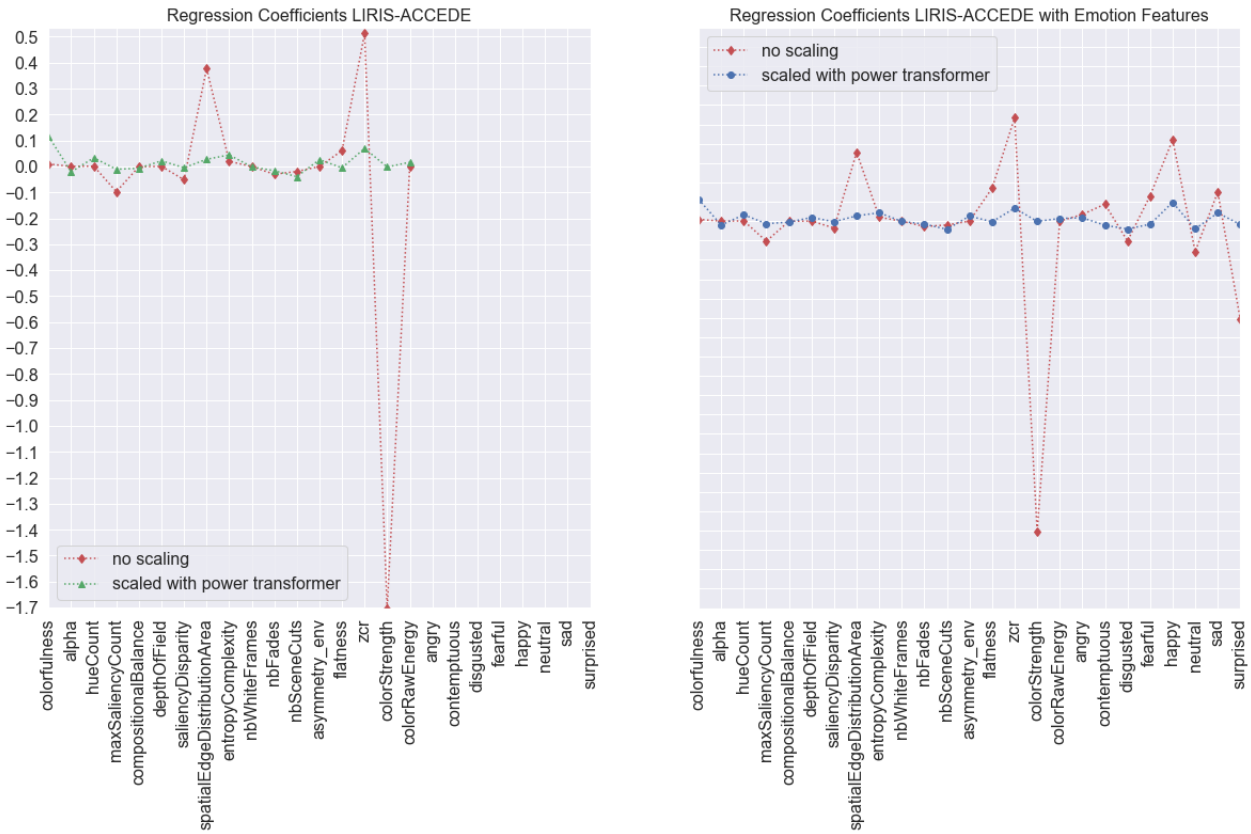

Figure A.6: Valence coefficients for linear regression.

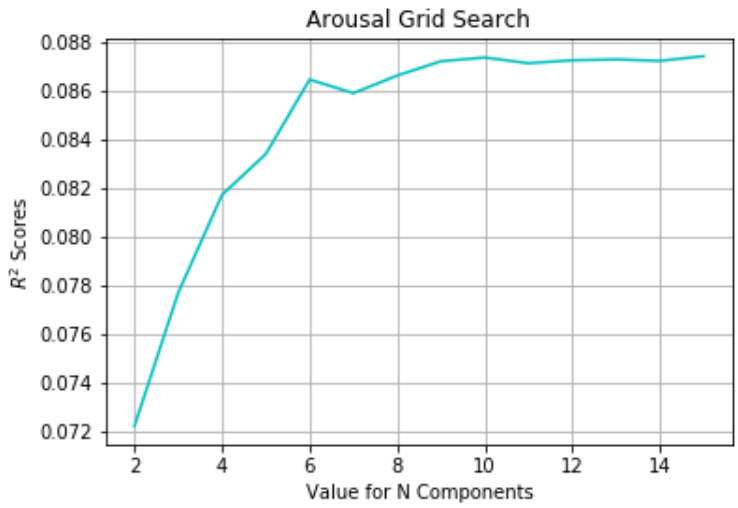

Figure A.7: Grid Search for PLS regression accounting arousal with emotions.

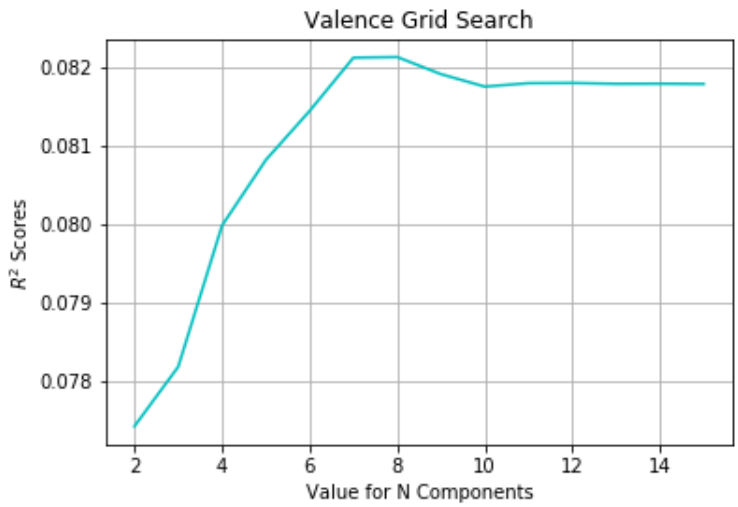

Figure A.8: Grid Search for PLS regression accounting valence with emotions.