Classification of valence using facial expressions of TV-viewers

Master’s Thesis

Yorick H. Holkamp
Classification of valence using facial expressions of TV-viewers

THESIS

submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Yorick H. Holkamp
born in Hellevoetsluis
Abstract

Emotion has been shown to have a large impact on our interactions with people and devices. In our daily lives, however, these emotions are not taken into account when working with our computers and other machines. If our devices could pick up on social cues, for instance in relation to disinterest, the usability of various systems could be improved.

Current software allows us to detect specific movements in people’s faces from video recordings. Using these movements, facial expressions can be linked to specific emotions, allowing for the incorporation of this information in various systems. One application would be to allow a TV to monitor its viewer, suggesting alternative videos to watch when negative emotions are shown.

An often used system to describe these specific facial muscle movements is the Facial Action Coding System (FACS). Despite the widespread use of this method, little research has been conducted on the use of FACS measurements to classify viewer emotion of entire videos. In this thesis we evaluated whether it is possible to use FACS measurements to perform classification on emotional labels in real-world environments.

To assess the possibility of this application, we conducted a wide range of experiments. We selected an existing method that uses a public dataset of naturally occurring emotions and reproduced this method. Additionally, we developed our own, alternative method. In a novel comparison we evaluated the performance of both methods on three different datasets, selected to cover a range of demographics and experimental settings (highly controlled to near-living-room conditions).

Furthermore we evaluated the inclusion of the TV viewer’s head orientation. This proved to be beneficial for two datasets. One of the datasets used in our work provided access to heart rate data of the subjects. Based on this data, we included the subject’s heart rate and other derived features. We found that this improved performance when training using the history of a specific person.

Finally we performed a novel experiment in which we asked a crowd of laymen to annotate videos from each of the three datasets. This multi-dataset evaluation provided us with a reference of how well humans were able to detect the emotion experienced by the subjects using their facial expressions, allowing for a direct comparison with automatic classification methods.
Overall we found that (1) using different data processing and aggregation, classification performance can improve and (2) that human annotation of emotional responses offers a way to compare classification difficulty between datasets and performance between classification methods.

Thesis Committee:

Chair: Prof.dr.ir. G.J.P.M. Houben, Faculty of EEMCS, TU Delft
University supervisor: Dr. C. Hauff, Faculty of EEMCS, TU Delft
Company supervisor: Dr. J.G.M. Schavemaker, Media and Network Services, TNO
Committee Member: Dr. H.S. Hung, Faculty of EEMCS, TU Delft
Preface

This report is the final result of my Master’s thesis and the conclusion of my Computer Science study at Delft University of Technology. During my thesis, I had the opportunity to conduct my research at the Media Networks & Services department of TNO. Originally starting as research into recommender systems, the focus soon changed to the recognition of emotion in TV viewers and has proven to be both challenging and educational.

I would like to take this opportunity to thank a number of people who helped me during my research. First of all my supervisors and colleagues at TNO have been a great help. John Schavemaker provided a steady source of ideas and valuable feedback throughout the project. In the first half of my project, Joost de Wit helped my research take shape, while Erik Boertjes helped improve the final result with feedback from a different perspective. At the university, Claudia Hauff guided my project to a successful end with critical questions and feedback whenever possible.

During my research I discussed my work with many people during the bi-weekly student meetings, at the university and at TNO, I would like to thank you all. Special thanks to the authors of the MAHNOB-HCI, BINED and MIME datasets. Without their work to collect these datasets and their permission to use these datasets, I would not have been able to conduct this research. Thanks to Veerle for her critical questions, love and support along the way. Finally, thanks to my parents for their everlasting support during my entire study.

Yorick H. Holkamp
Delft, the Netherlands
August 27, 2014
# Contents

Preface ........................................ iii

Contents ....................................... v

List of Figures ................................ vii

1 Introduction ................................ 1
   1.1 Research questions ........................ 3
   1.2 Contributions ............................. 4
   1.3 Outline .................................. 5

2 Background & Related work ............... 7
   2.1 Affective computing ......................... 7
      2.1.1 Emotions .............................. 8
      2.1.2 Facial expressions ....................... 9
      2.1.3 Machine learning classifiers ............ 11
   2.2 Emotion classification ....................... 12
      2.2.1 Facial expressions ....................... 12
      2.2.2 Multi-modal emotion classification .... 14
      2.2.3 Data processing ........................ 15
      2.2.4 Human-machine comparisons .............. 16

3 Datasets .................................... 19
   3.1 MAHNOB-HCI ................................ 20
   3.2 BINED .................................... 21
   3.3 MIME ..................................... 22
   3.4 Data selection ............................. 22

4 Automatic and Human Labeling .......... 25
   4.1 Automatic classification .................. 25
      4.1.1 Pipeline overview ..................... 26
      4.1.2 Onset/Offset Counting ................ 27
      4.1.3 OOC implementation .................... 28
      4.1.4 Facial expression intensity method ... 29
List of Figures

1.1 Example use case for an emotion recognition system. .................................. 2
2.1 The 2-dimensional pleasure-arousal space, as presented by Feldman Barrett and Russell [1998]. .......................................................... 8
2.2 Example Action Units from the Cohn-Kanade dataset, by Zhang et al. [2008]. 10
3.1 Example dataset collection process. .............................................................. 20
3.2 Sample frame from the MAHNOB-HCI dataset. [Soleymani et al., 2012] 21
3.3 Sample frame and photo of the experimental setup from the BINED dataset. [Sneddon et al., 2012] ............................................................. 22
3.4 MIME experiments. .................................................................................. 23
4.1 Overview of an automatic emotion classification pipeline. ...................... 26
4.2 Video annotation interface used in the human annotation experiment. ... 31
5.1 Classification rates of human annotators, random and majority class classifiers on the three datasets. ......................................................... 41
5.2 Classification rates for human annotators and the intensity-based method on the three datasets considered. ............................................... 42
Chapter 1

Introduction

Emotions play a big role in our daily lives. During the course of a day, we experience a wide range of emotions, both in reaction to all kinds of actions and in conversations with others. These emotions affect decision making and have an impact on our interactions with our environment and peers.

Meanwhile, once we change from our interaction with people to interacting with machines, none of these emotions are taken into account. Our smart phone will not pick up on social clues such as an angry tap on a certain application, nor will our TV detect the angry glances we might give it. If our devices were able to pick up on these social clues, we might experience our interaction with them as more natural and pleasant.

The recommendation engines used by services such as Netflix or Amazon try to reason whether we enjoyed a certain TV show or product based on our behavior. These systems rely on indirect information, i.e. stopping a movie halfway in, or explicit direct feedback, such as a rating. At the same time, television sets have been improved to the point where nowadays many of new devices being sold have an embedded video camera, allowing it to look at its viewer. Nowadays, several commercial solutions exist that are capable of recognizing facial expressions, such as Microsoft’s Kinect for Xbox One, making it feasible to incorporate this facial expression data in these systems.

The importance of facial expressions in conveying our emotion as described by Mehrabian [1968] fits well with the early work on facial expressions by Darwin [1872], who also stressed the link between facial expressions and emotion. A second important finding was made by Ekman et al. [1987], who later found that many emotions are universally recognized between cultures, using the same set of facial expressions. This indicates that a system with the capability of recognizing facial expressions would be able to work with a wide range of people. Combining the relatively old concept of video recommendation with the new found ability of our television sets to observe its viewer enables us to gather facial expression information. By interpreting this data, as done in the field of social signal processing, it may be possible to improve the ease of use of these devices [Arapakis et al., 2009b, Soleymani and Pantic, 2013].

In Figure 1.1 we show an example system which incorporates these facial expression recognition techniques. Here, we show a viewer who watches videos on a television in her living room while being recorded with a video camera. This video data can then be converted. In this step facial expressions shown by the viewer are extracted,
Introduction

which results in a stream of processed data. This processed data is then sent to the
data interpretation component, along with a set of previously observed behavior; the
training data. In this component, relevant aspects of the data are selected, manipulated
and used in a machine learning algorithm. Based on patterns found in the training data,
this algorithm then relates the observed behavior to a specific emotional label. Finally,
this emotional label can be used to refine video recommendations, which the TV could
present to the viewer.

One way of refining these recommendations would be to compare the emotion of
the viewer with the expected emotion of a certain movie; if the viewer is looking bored
during a comedy show she might be interested in receiving a recommendation for an
alternative program. To the best of our knowledge, no such systems are available yet.
To facilitate this kind of functionality, we look at existing affective computing research,
where the incorporation of emotion in computer systems is a core research topic.

To facilitate research into these kinds of applications, various datasets with ex-
ample data exist. These datasets contain videos that show subjects watching a video
clip, after which these subjects reported the emotion they experienced during the view-
ing of this video fragment. Using these examples, we can try to discover patterns in
the behavior of people who describe having felt a certain emotion, which can then be
used to predict these felt emotions in different situations. The development of accurate
emotion detection methods are the first step towards the development of successful
affective computing systems. Without an accurate method of detecting emotion, any
consecutive interpretation becomes irrelevant.

While the reported emotions can cover a wide range of possibilities, we will pri-
marily focus on the classification of valence. Russell [2003] described that valence
is an essential aspect of any emotional response. This dimension covers the differ-
ence between positive and negative experiences, allowing for a powerful distinction of emotions in various applications. The use of valence has been seen in analysis of viewer responses to advertisements for market research purposes [Garbas et al., 2013], software usability testing [Hazlett and Benedek, 2007] and affective video recommendations [McDuff et al., 2010]. Furthermore we argue that positive emotions, such as happiness, are more likely to be experienced when watching an interesting video fragment than negative emotions, such as boredom. This brings us to our core challenge; how can we determine the emotional valence a TV viewer experiences using information obtained from video recordings of this person?

In order to answer this question, we have defined several research questions which we will address in this work. A detailed description of these questions is provided in Section 1.1. Next we will describe the contributions of our work in Section 1.2. Finally we describe the outline of this thesis in Section 1.3.

1.1 Research questions

When we consider the scenario outlined in Figure 1.1, we can see that there are several aspects to an emotion-aware system. In this work we focus on the data interpretation block shown in the figure, which handles the problem of interpreting the pre-processed sensor data and relating this information to an emotional label. This has led us to one central question which will be considered throughout this work. In order to work towards an answer for this question, we have decomposed it into five research questions, each providing answers to a part of the challenge. Our main question is as follows:

Main research question How can the classification of the emotion of people watching a video fragment be applied in a real-world setting when only using video recordings depicting the viewer?

Many methods are described in literature that are able to perform emotion classification, as we will see in our background chapter. One pattern we have seen in the encountered works is that several of these methods rely on prior knowledge of the behavior of one specific person for the system to be able to predict the user's emotion in new situations. [Tkalcic et al., 2011, Koelstra and Patras, 2012] In a real-world application, such information would not be available. Take for instance the scenario where a consumer purchases a new television that is capable of emotion recognition, in that case we would not be interested in waiting a long period of time for the system to get acquainted with our use of facial expressions, before it is able to assist us in whatever way possible. With the following research question we try to determine whether these methods are suitable for use where no prior information about a specific user is known.

Research question 1 When the system has not seen a specific user before and is not familiar with their behavior, to what extent can methods found in literature be successfully applied?

Now that we have seen methods as described in literature, we want to explore the effects of changing components of these methods will have on the classification performance. In Figure 1.1 we have seen that the actual interpretation block consists of three main steps; data selection, manipulation and machine learning. Our investigation
will focus on whether improvements can be made over existing solutions by changing the way the sensor data is processed and aggregated.

**Research question 2** What is the influence of alternative data pre-processing and feature vector construction methods on classification performance?

Various datasets of affective displays are available for research in the field of affective and emotional displays. [Kanade et al., 2000, Pantic et al., 2005, Lucey et al., 2010, Sneddon et al., 2012, Soleymani et al., 2012] There are however, limitations to the similarity between these datasets and the real world. Some datasets, for example, rely on subjects posing with a certain facial expression, which may be more expressive than spontaneously occurring expressions [Sebe et al., 2007]. Moreover, some datasets rely on highly controlled environments with specific lightning and camera positions. The choice of dataset may influence the performance of these methods, which led us to the next question.

**Research question 3** How well can facial expression-based emotion classification methods be used in uncontrolled environments, such as living rooms around the world?

The answers to the previous research questions might produce methods to perform emotion classification at certain performance levels, evaluated in more or less controlled environments. While this is a step forward, we are also interested in obtaining a target performance rate for these systems to work towards, providing a clear benchmark. Using such a benchmark, we would also be able to determine whether or not the performance has reached or is close to levels deemed acceptable.

**Research question 4** How can we utilize people to determine a benchmark level of valence classification accuracy for emotion classification methods?

### 1.2 Contributions

With this work we contribute to the field of affective computing. Below we provide a brief summary of these contributions, following the order of the research questions as presented in Section 1.1.

1. **Evaluation and improvement of affect labeling methods in cold-start situations.** We evaluated the performance of the method presented by Koelstra and Patras [2012]. This method was selected for its use of a public dataset and focus on labeling of affective responses to videos. In our evaluation we also included a novel method based on principles described in literature. These methods were trained on a training set of subjects, after which the testing was performed using a subject not included in this training set.

2. **Assessment of the impact of different pre-processing and feature vector construction methods.** We present a novel method that uses different pre-processing and feature vector construction methods. This novel method was compared with the method by Koelstra and Patras and improved on the performance of this method in most situations.
3. **Evaluation of methods on multiple datasets.** We presented a novel approach in which we evaluated the performance of classification methods on multiple datasets of naturally occurring emotions. The datasets were chosen to cover a range of different demographics, stimulus videos and experimental environments.

4. **Human labeling experiments on affective datasets and comparisons with automatic methods.** We conducted a human annotation experiment with a crowd of laymen who annotated videos from three different datasets with affective ratings. Using the results obtained in these experiments, we compared the performance of human and automatic labeling methods in a broad comparison on datasets not yet evaluated in this manner.

### 1.3 Outline

This work is structured as follows. First background knowledge and an overview of related work will be provided in Chapter 2. With the background of our work described and related work identified, we will introduce the datasets that have been used in this work in Chapter 3. Then we describe the classification methods used in Chapter 4. Next the evaluation of these methods is described in Chapter 5, where the performance figures are also presented. The results obtained in our experiments are discussed in Chapter 6. After this we conclude our work in Chapter 7 with an evaluation of our research questions as well as the identification of future work. For the convenience of the reader a glossary of used abbreviations and concepts is provided in Appendix A.
Chapter 2

Background & Related work

In order to place the work conducted in this thesis in the proper context, we provide background information as well as related work in this chapter. We will start off with an introduction to the topic of affective computing, which is the research area that covers the use of emotion in computer systems. Here we also provide background information on the topic of emotion and facial expressions, which are both essential components for our work. Next we continue with an overview of methods and techniques used on the topic of emotion classification from videos, followed by a summary of the approaches taken and an overview of comparisons between human and machine classifier accuracy.

2.1 Affective computing

Affective computing is a field that tries to “assign computers the human-like capabilities of observation, interpretation and generation of affect features” [Tao and Tan, 2005]. To do so, the systems and methods developed in this field typically require knowledge from computer science, psychology, which provides the knowledge into emotion and affect, and cognitive science, which focuses on the cognition aspects.

The modern day interpretation of affective computing gained traction after the pioneer paper by Picard [1995]. In her work, she described a future where computers are able to interpret the emotion of its user, adapt its behavior according to this emotion and even portray emotion. By adapting to the user’s emotion, the software could then be made to appear more creative and less predictable, which could lead to positive outcomes.

While we are not currently researching options to have our computers show emotions, we are interested in determining whether we can develop a method to interpret the user’s emotion. These interpretations could then be used to adapt the behavior of our software to this emotion. An example of such a system, which relies on such a component, has been presented by Klein et al. [2002]. In their work they describe a computer that adapts to the frustration of its user and they found that their subjects found this to significantly improve their user experience.

As affective computing includes both computer science and psychology or cognitive sciences, it is an interdisciplinary field. For us this means that we will benefit from a wide range of knowledge on topics ranging from emotion to machine learn-
2.1 Affective computing

Background & Related work

Before developing a computer system that takes into account human emotion, it is necessary to have some understanding of how to classify emotion and to have a basic understanding of how expressions of emotion might differ between people. Once we have observed this, we also provide a brief overview of commonly used machine learning algorithms in affective computing systems.

2.1.1 Emotions

One aspect that plays a large role in the affective computing field is the concept of emotion, which, when experienced, is often referred to as “affect” or “affect(ive) displays”. Kleinginna Jr and Kleinginna [1981] showed that the construction of a single definition of emotion, even from a list of hundreds of definitions, appears to be a topic worthy of its own research. While providing a definition of the concept might be difficult, several ways to describe emotional states have been developed. Ekman et al. [1987] found that a number of emotions are recognized across different cultures. The six original emotions discovered to be universally recognized, based on facial expressions, are referred to as basic emotions. These emotions are anger, disgust, fear, sadness, happiness and surprise. In later work [Ekman, 1992], several other emotions were included as well, e.g. contempt.

Alternatives to Ekman’s method of emotion classification are also in use. One example of this is the pleasure-arousal-dominance (PAD) dimensional model, which does not rely on discrete labels. In this method, a 2- or 3-dimensional space is used to describe emotion. A 2-dimensional example of this is shown in Figure 2.1, as first presented by Russell [1980]. The two common dimensions used in these spaces are valence (positive or negativity of the emotion) and arousal (the level of reaction). Here, sadness would have a negative valence and a low arousal value, whereas happiness would have a positive valence and neutral to high arousal value. Additionally, Mehrabian [1980] described a variation of this approach which also includes a third “dominance” dimension, which describes whether an emotion is dominant or controlling. The argument for the additional dimension is that while fear and anger are both negative valence and high arousal emotions, anger is a more dominant emotion than fear, which is more submissive.

![Figure 2.1: The 2-dimensional pleasure-arousal space, as presented by Feldman Barrett and Russell [1998].](image-url)
2.1 Affective computing

Demographic differences

Although we have seen there is a strong indication that there is a set of universal basic emotions, as described by Ekman [1992], there may be demographic differences in how these emotions are expressed. For instance, it is evident that not all individuals express their emotions at the same intensity. For example, when two individuals experience the same levels of happiness, they will likely show different smiles.

Another contributing factor to these differences is the age of the subject. Malatesta et al. [1987] conducted experiments where subjects of varying age had their facial expressions recorded during an emotion elicitation experiment and a group of observers were then asked to determine which emotions were expressed. Their experiments suggested that the facial expressions of elderly subjects are more difficult to decode than those of younger subjects. There are also indications that elderly people might have a tendency to mask their true emotion with smiles, confusing the interpretation as shown by Keltner [1996].

At the same time, the interpretation of emotion is influenced by many factors. Hess et al. [2000] showed that the interpretation of facial emotion displays is influenced by the gender and ethnicity of the person displaying these expressions. Similarly, Matsumoto [1993] showed that the ethnicity of the observer has a significant impact on the interpretation of emotion intensity.

These differences in the displays of emotion between people of different age and ethnicity are shown to complicate the interpretation of emotion for human observers. When using facial expressions for affective computing purposes, it might therefore be necessary to account for these differences in order to properly determine the emotion displayed by the user.

2.1.2 Facial expressions

Humans make extensive use of nonverbal communication with facial expressions playing the major role as shown by Ekman et al. [1987]. In order to be able to objectively quantify facial movements, Ekman and Friesen [1978] developed the Facial Action Coding System (FACS). In this system careful descriptions are provided to quantify specific facial movements, which are called Action Units (AUs). FACS defines 44 unique AUs, 30 of which relate to specific muscle contractions, 12 for the upper part and 18 for the lower part of the face. Some examples of these AUs are shown in Figure 2.2. Example studies using this method include differentiating between displays of genuine and simulated pain [Craig et al., 1991] and detecting when a person might be telling a lie [Ekman, 2009].

FACS does however have its limitations. While Ekman and Friesen described into great detail how certain facial expression features can be recognized and classified, FACS does not include a mapping from these AUs to emotions. Two methods exist to bridge this gap: Emotional FACS (EMFACS) [Friesen and Ekman, 1983] and FACS Affect Interpretation Database (FACSAID) [P. Ekman and Hager, 2003]. These methods only consider emotion-related AUs and relate AUs to discrete emotions. The main difference between these two methods is that EMFACS only considers basic emotions while FACSAID includes a wide range of affective states, including basic emotions,
2.1 Affective computing

**Background & Related work**

Figure 2.2: Example Action Units from the Cohn-Kanade dataset, by Zhang et al. [2008].

<table>
<thead>
<tr>
<th>Happiness</th>
<th>Fear</th>
<th>Sadness</th>
<th>Surprise</th>
<th>Anger</th>
<th>Disgust</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1+2+4</td>
<td>1+4</td>
<td>1+2+5+2627</td>
<td>4+5+7+17+2324</td>
<td>9110+17</td>
</tr>
<tr>
<td>6+12</td>
<td>1+2+4+5</td>
<td>1+4+1115</td>
<td>1+2+5</td>
<td>4+5+7+2324</td>
<td>9110+16+2526</td>
</tr>
</tbody>
</table>

Table 2.1: Example of combinations of AUs in relation to the six basic emotions, as described by EMFACS, “+” and “|” represent AND and OR, respectively. [Koelstra and Patras, 2012]

and attempts to cater for a wide range of psychological states. An example of AU-emotion-relations is shown in Table 2.1.

Both methods are however limited in that they only relate AUs to discrete emotions, without any link to the 2- or 3-dimensional models of emotion. Limited work is available on the relation between AUs and dimensional models of emotion as was described by Koelstra and Patras [2012]. It might however be possible to relate these emotions to arousal and valence values using the semantic space for emotions outlined by Scherer [2005].

**Automatic detection of Action Units**

In addition to the limitations of the FACS method in relation to dimensional models of emotion, there is another challenge. FACS was originally designed to allow human annotators to analyze videos frame by frame. The annotators would then manually describe which of the Action Units were activated and at what level. While Ekman and Friesen showed that this allows for very high accurate ratings with high agreement among raters, it is very time consuming.

To allow for more practical applications as well as cheaper and faster experimenting with FACS, many researchers have developed methods of automatically detecting Action Units using software [Donato et al., 1999, Tian et al., 2001]. These methods either work on specific photo of a face or even sequences of images (videos). The de-
Background & Related work

2.1 Affective computing

Development of these systems has been supported by the creation of large datasets such as the Cohn-Kanade dataset [Kanade et al., 2000] and the MMI database [Pantic et al., 2005]. These dataset were specifically developed for the purpose of automated facial expression recognition and have, in the case of the Cohn-Kanade dataset, subsequent updates [Lucey et al., 2010].

Nowadays several applications capable of face detection and automatic facial expression recognition exist, both commercial and academic. One such academic solution is CERT [Littlewort et al., 2011], whose authors now produce a commercial solution as Emotient\(^1\). A second commercial solution called Affdex is produced by the MIT spin-off Affectiva [Picard, 2011]. The final solution we have found is the Noldus FaceReader [Den Uyl and Van Kuilenburg, 2005].

In addition to the recognition of different Action Units, these applications often also provide an interpretation of the observed behavior in the form of a \([0, 1]\) rating for six basic emotions, allowing for a more direct interpretation of the observed behavior.

2.1.3 Machine learning classifiers

To link the user behavior observed using a range of sensors to specific classes, for instance “happy” and “bored”, affective applications frequently use machine learning classifiers. These algorithms are able to recognize patterns in the data using techniques that are different for each algorithm. Typically the algorithms rely on an n-dimensional feature vector of values that represent each instance of an object.

In a simple case, a feature vector could consist of the number of seconds a user looked at a specific photo. When combining this information with the rating of a user as a so-called ground truth, a machine learning algorithm might be able to discover a pattern between the feature vectors and their respective ground truths. Using this information it can then be used to predict the in the data and use this to predict the class for new instances.

An overview of commonly used machine learning algorithms in affective computing literature was presented by Scherer et al. [2010]. In their work they presented a number of algorithms which they describe as being “best known” for classification purposes in the field. The two algorithms that are relevant for our work are Support Vector Machines (SVM) and k-Nearest Neighbor (kNN) algorithms. The list mentioned by Scherer et al. is however not exhaustive. For example, Nwe et al. [2003] described that Hidden Markov Models (HMM) are frequently used for emotion analysis in speech. For each of these algorithms we provide a brief description.

**k-Nearest Neighbor** Calculates the similarity between the values for the current object and previously seen examples, resulting in a number of neighbors. The chosen class is then calculated using a majority vote of the n-nearest neighbors.

**Support Vector Machine** Draws a plane that separates the feature vectors of one class from the feature vectors that belong to a different class with the maximum margin.

**Hidden Markov Model** Describes the concept under consideration as a Markov process, where the different states of the Markov model cannot be observed. This

---

\(^1\)http://www.emotient.com/products
algorithm is in particular well-suited for problems where a specific sequence of values should be recognized, such as speech recognition.

Among others, Rani et al. [2006] evaluated the performance of SVM and k-NN algorithms in an emotion classification setting. In their work, they found that Support Vector Machine (SVM) consistently produced better results than the kNN and other algorithms. The authors described that the use of kNN can still be beneficial in some cases. In particular, when the machine learning algorithm will frequently have to be re-trained there may be a significant speed advantage to using the kNN algorithm, as it requires no training, whereas this may take several minutes up to several hours for large datasets when using a SVM algorithm.. Based on what we have seen, it depends on the application in mind which classification algorithm works best.

2.2 Emotion classification

We know that correctly recognizing the emotion of a person is not an easy task, even for the human observer. In our daily lives we will take into account various sources of information when estimating a person’s emotional state, these sources are referred to as modalities. We can listen to the sound of a person’s voice, consider their posture or movements and look at facial expressions. When trying to determine the emotion of a person using software, depending on the environment and available sensors, these modalities can also be used.

Now that we have glimpsed at the foundations of affective computing, we may have a more detailed look at some of the methods that have been devised to perform emotion classification. In the following sections, we will first discuss some of the existing methods that rely on the use of only facial expressions. Next we will consider the multi-modal approaches, which take into account a broader range of data. These methods can be more similar to how we estimate emotion as humans or go much further, by using heart rates and other data that is not available in our day to day lives. After we have seen these methods, we present a short overview of some of the data processing steps used in the published works. Finally we take a look at some of the comparisons that have been published between human and machine learning classifiers, as these might provide a reference level of performance helping us identify the practical usability of these methods.

2.2.1 Facial expressions

The use of facial expressions in affective computing systems has been evaluated by many researchers. In particular, its use in the context of content retrieval systems has been evaluated in various works. One such example was published by Arapakis et al. [2009b], who included facial expressions reported by the eMotion software to estimate the quality of produced recommendations in a document recommender system. In their system, a SVM was trained to predict the user’s liking based on the facial expressions detected in the few seconds after a document was shown.

In a similar setup, Tkalcic et al. [2011] evaluated the use of facial expression information when retrieving implicit feedback for a photo recommender system. In a
comparison between a more traditional baseline recommender system, which considered the time watched and genre of the photo, their recommender system that also considered the PAD dimensions obtained a higher performance.

In addition to these systems aimed directly at recommender systems, other applications have been evaluated as well. One such evaluation was presented by McDuff et al. [2010], who presented a purely facial expression-based method to estimate valence ratings of a video. In their work, they used a dataset consisting of subjects who watched several movie fragments known to elicit specific emotions. While the subjects watched these videos, they provided a valence rating by positioning a slider such that it matched their emotional state at that specific time. Moreover, the facial expressions recorded for these subjects were manually annotated by two certified FACS annotators, providing their method with high quality facial expression measurements. Using the per-second Action Unit activity provided by the annotators they then compared the ability of different machine learning algorithms to predict whether the valence of the subject was positive, neutral (occurred infrequently) or negative during the considered second. They evaluated the performance of these methods using increasingly general training methods, with the most general option being cross-validation where the sessions for one specific movie were left out. In this scenario, the accuracy of the SVM method was around 48%.

A similar method with different end-goal was presented by Joho et al. [2009], who used facial expressions to detect personal highlights in videos by observing the viewer. In order to collect the necessary data for these experiments they asked their subjects to first watch a video and afterwards annotate the parts of the video they considered to be their personal highlights. The information from their facial expression recognition software, which used a custom system similar to FACS, was then used to construct various features, which were eventually used to perform binary classification on the video frames. In particular they evaluated the use of features that reflected how pronounced the facial expressions of the subject are, the rate of change of these emotions and a feature-level fusion of these two.

A more recent evaluating the classification of emotion using Action Units was conducted by Senechal et al. [2014]. In their work they compared the performance of two automatic classification methods; one that received features extracted from the video frame being classified and a second method that received Action Unit scores. These two methods were then evaluated using three different datasets containing photos of people displaying certain emotions. In this evaluation, they found that the classifier using AU scores obtained a higher accuracy than the method using video-based features.

While these methods present the use of facial expressions a similar context, these methods are using only limited information. Rather than observing their user for a longer period of time, shortcuts are taken. They either rely on the use of facial expressions observed during a single video frame [McDuff et al., 2010] or monitor the user during the short period of time after the user viewed a photo or document [Arakakis et al., 2009a, Tkalcic et al., 2011]. Both of these methods greatly reduce the noise that is contained in the measurements, which simplifies the classification process.

When recommending items that require the attention of the user over a longer period of time, such as video fragments, it is however not possible to limit the observation to just a few seconds. There exist methods that do monitor the user over a longer period of time, such as Koelstra and Patras [2012]. These methods do however also
incorporate more than just facial expressions and will be discussed in the next section.

### 2.2.2 Multi-modal emotion classification

In some affective applications, more modalities might be available than just facial expressions. For instance, in a lab environment it is possible to measure a subject’s heart rate, skin conductance and brain activity and use this information for our emotion recognition task [Koelstra and Patras, 2012, Koelstra et al., 2012, Janssen et al., 2013]. Currently, each of these modalities require the use of sensors that are physically attached to a person’s body, we therefore refer to these methods as “invasive”. On the other hand, non-invasive alternatives do not require such attachment but rather function from a distance. Examples include the user’s voice or their facial expression.

One particular non-invasive modality is to analyze the speech of a subject. Such a method was presented by Nwe et al. [2003], who employed a Hidden Markov Model classifier using feature vectors constructed using a specific spectral analysis method to represent the speech. They found that using spoken text, they were able to classify speech samples into one of six basic emotions with an average accuracy of 78%. This is significantly higher than would be expected from a randomized system. While in their work this was the only modality considered, it serves as an example of what might be possible using sound.

However, many of the methods found in literature rely on combinations of modalities. In order to combine these modalities, different options are available. The first option is to combine the features obtained from both modalities into a single feature vector and use a single machine learning algorithm to perform classification using these vectors. This is called feature-level fusion. Alternatively, decision-level fusion can be used. In this case, two separate feature vectors and machine learners are set up, which independently predict the proper label. After this has been done, the final label is chosen using the predictions of these two classifiers. The ways these labels can be combined into a final classification are plentiful, as was shown by Ruta and Gabrys [2000].

An early multi-modal approach that evaluated both feature- and decision-level fusion is presented by Busso et al. [2004], in their work they compared emotion detection using video and audio based features. Their corpus consisted of a dataset of videos recorded of an actress, who read a number of sentences, each with a target of one of three emotions or a neutral state. From these recordings, facial expressions and acoustic information were extracted. Using these features, they trained SVM classifiers using combinations of these two modalities using both feature-level and decision-level fusion. In their evaluation they found that the use of both audio and video modalities proved to be beneficial for the classification accuracy. They found that the different fusion methods resulted in different classification rates for each specific emotion, leading to the conclusion that the best fusion method depends on the application used.

In more recent work, the fusion of invasive and non-invasive modalities is frequently evaluated. One such evaluation was conducted by Koelstra and Patras [2012], who evaluated the use of both facial expressions and EEG data. Both feature- and decision-level fusion of classifiers for these two modalities were considered. Using these features, they performed binary classification on valence, arousal and control targets using a Gaussian Naive Bayes (GNB) classifier. In their work they found that
the use of EEG data resulted in higher accuracy and F1 scores than the use of only facial expressions. Moreover, a feature-level fusion resulted in higher performance figures than the individual modalities, however, little difference was found between the different fusion methods.

Janssen et al. [2013] presented a method for emotion classification that includes even more modalities using a SVM classifier. In their work they included features extracted from video and audio as well as physical sensors, such as skin conductivity, skin temperature and features derived from the subject’s heart rate. They found that the combination of physiological features with other modalities resulted in the best performance, with just audio-based features performing worst, followed by video-based features. The dataset used in their work was however not directly comparable to the dataset considered in the previous work by Koelstra and Patras. Rather than observing people who are watching a video, they asked their 17 subjects to describe situations in which they experienced specific emotions. The use of this so-called autobiographical recollection is known to induce strong emotions in the subjects [Levenson et al., 1991] and requires them to speak, resulting in audio recordings. During these sessions the subjects were recorded using audio, video and various physiological features.

We have now seen that facial expression recognition is not a simple task, even for human observers, the method used to train a classifier may also impact the performance of the method. Tkalčič et al. [2013] conducted an experiment where they compared the and generic discuss the performance of several machine learning classifiers when faced algorithms using two datasets, one with “weak” emotional displays and one with clearly posed emotions. The authors find that the detection accuracy drops significantly when a weak dataset is being used, stressing the requirement of quality training data.

2.2.3 Data processing

The observation of subjects while they watch video fragments will typically result in sensor values at discrete intervals during the video playback. On the other hand, most machine learning algorithms rely on a single n-dimensional vector to characterize such a session. So to be able to feed the measurements to a machine learning classifier, it is necessary to aggregate the measurements over time to obtain a limited number of values. The specific aggregation method used differs between the approaches described in literature.

The simplest aggregation method possible is to count the number of times a sensor value passes a certain threshold. In the context of facial action data, this could effectively correspond to the number of times a subject raised their eyebrows during video playback [Koelstra and Patras, 2012].

A different approach was proposed by Yeasin et al. [2004], who presented a method to derive the level of interest of a person using their facial expressions. This level of interest can then be used as a new measurement value and averaged to obtain a single value for use in a classification context.

A similar time-continuous approach was created by Joho et al. [2009], who derived two aggregated features from their facial expression data. First of all they observed the rate of change of facial actions and second they observed how pronounced the subject’s facial action display was. These metrics resulted in functions that were continuous in
time but could be used to determine, for instance, the average value throughout the video, where peaks could be found or where those were absent.

In addition to these simple aggregation methods, a method used by Tkalcic et al. [2011], Janssen et al. [2013] is to calculate the first two statistical moments of the feature values: which are the average and the variance. In particular the variance was reported to be frequently chosen as distinctive by the machine learning algorithm in these works.

Typically, once the feature vectors have been created, post-processing is applied. One step that is often performed is normalization. In this step the vectors are normalized using some method in order to avoid features with large numeric values from overshadowing the features with smaller values [Hsu et al., 2003]. Common options are to linearly scale the features to a limited range, such as $[0, 1]$ or $[-1, 1]$. Alternatively, the vectors can be scaled to unit norm, such that the length of the vector is equal to 1, either measured using the Manhattan (L1) or Euclidean norm (L2 norm) [Burges, 1998].

A second step that can then be considered is to perform feature elimination on the feature vectors. By reducing the number of features included in the feature vector, the classification speed can be improved. Furthermore, when the number of features is very large compared to the number of instances, this can avoid the system from overfitting the training data. One popular option is the Recursive Feature Elimination (RFE) method presented by Guyon et al. [2002]. In this method a SVM is trained using a set of features, recursively the features assigned the lowest weight are removed and the SVM is re-trained until a predefined number of remaining features has been reached.

An alternative method, used by Koelstra and Patras [2012], is Independent Component Analysis (ICA). The goal of this method is to create a linear representation of the data, such that the components are as statistically independent possible [Hyvärinen and Oja, 2000]. Koelstra and Patras, who compared RFE and ICA, reported the best results using RFE.

We have now seen that there are different aggregation methods that can be applied to transform sensor values over time into a limited number of values. The best aggregation method however depends on the application and type of available data. Furthermore the importance of feature vector normalization has become clear, as well as the possible added value of applying feature elimination.

### 2.2.4 Human-machine comparisons

Even though automatic emotion recognition methods have improved, researchers such as Calvo and D’Mello [2010] posit that the accuracy of automated affect recognition is currently not yet high enough for use in real-world systems. The question of when the accuracy will be high enough for practical applications remains a complicated one. One way to determine whether this is the case might be to compare these systems with human annotators with see how they compare.

There have been several researchers who conducted such comparisons between the performance of human and machine classification. Frequently, it is discovered that the automatic classification manages to perform better than human observers. For instance, Nwe et al. [2003] compared the performance of human and software in speech-
based emotion classification. They used a corpus consisting of audio fragments from
an actor who spoke lines of text in an emotional voice. In their study, the software
outperformed the recognition rate of their human annotators. In a similar experiment,
Esparza et al. [2012] found that human annotators were not able to match the accuracy
seen by automatic classification methods.

A similar study was conducted by Susskind et al. [2007], who compared facial
expression recognition by human and computer. The focus of their work was to deter-
mine whether a machine learning algorithm trained to recognize one particular emotion
would be able to obtain performance similar to human observers. The results of their
experiments show that the accuracy obtained by human and computer were of similar
magnitude. The corpus used in their experiment however consisted of photos of people
who posed a certain emotion, limiting the similarity to real-world situations.

More recently, an elaborate comparison was presented by Janssen et al. [2013],
who used audiovisual recordings of their subjects rather than photos. In their work they
compare the classification performance of two machine learning methods (described
in Section 2.2.1) with the classification performance of human observers. In addition,
they conducted two experiments with human annotators. These annotators observed
the video and/or audio footage, depending on the experiment they participated in. To
avoid the semantic meaning being interpreted by the observers, their subjects did not
speak the language spoken in the audio recordings. Overall they reported that their
machine learning methods outperform the human observer accuracy when provided
with the same audiovisual information in a controlled environment.

Now that we have seen multiple studies where the human classification accuracy
was compared with automatic systems, it is safe to say that the performance of auto-
matic systems might in fact be relatively high. In the studies we described, the machine
learning algorithms outperformed the performance of the human observers when both
provided with the same video or audio data collected in specific settings. Whether
this comparison holds for more flexible situations and different methods of emotion
elicitation in subjects is however unclear.
Chapter 3

Datasets

The use of emotion recognition from video recordings allows for many applications in software. In future systems, it would be possible to determine whether a user is frustrated by combining information regarding the way user is currently using an application and their facial expressions. After such emotions have been identified, the application could provide the user with additional support. Alternatively, the collected information could be used to provide recommendations tailored towards the user’s emotion.

In order to develop such applications, large amounts of data are required to train automatic classification systems to recognize patterns in the user behavior. Specific datasets tailored to suit these needs exist and were typically collected through experiments conducted in an academic setting. In these experiments, subjects were asked to watch a video clip, the stimulus. These stimuli are often chosen to induce specific emotions. After a video has been viewed, the subject is asked to describe the emotion they felt during the video, a so-called emotional self-assessment. Additionally, the subjects were recorded while they were viewing the stimulus videos. In this work, we will refer to the recording of a subject combined with their emotional assessment from one stimulus as a session of the subject. Typically these datasets consist of hundreds of sessions showing many subjects.

The different datasets that are available for emotion recognition vary in many ways. Main variations can be found for instance in the type of environment in which the experiment was conducted, ranging from laboratory to living room settings. Moreover, the demographics of the subjects and the stimulus videos that were used differ between datasets. The demographics range from undergraduate students to elderly people. Similarly, the stimulus videos range from fragments of regular television shows to explicit medical surgeries.

In our work we have used three different datasets, which were chosen to cover these variations. Each dataset will be described in detail in the following sections. The datasets consist of the raw data collected during the experiments, therefore minimal post-processing has been applied. This means that there may be sessions present which cannot be used. In these cases video footage might be missing, the subject might not have completed watching the entire experiment or the self-assessment rating might have been lost. To ensure we only work with sessions that have been successfully obtained, we have used a set of exclusion criteria, which we will present in our final
3.1 MAHNOB-HCI

The MAHNOB-HCI dataset, short for Multimodal Analysis of Human Nonverbal Behavior: Human Computer Interfaces, was presented and developed by Soleymani et al. [2012]. It is a dataset developed for the purpose of affect recognition and implicit tagging research, the latter is the concept of automatically assigning labels (tags) without manual interaction. The dataset contains recordings of 30 subjects, who each watched 20 stimulus videos in a lab environment. These subjects were of various nationalities, living in or nearby London, 17 of which are female and 13 male. The stimulus videos were obtained from popular movies as well as sites such as YouTube.com and shown in random order. Additionally neutral video clips were shown in between videos in order to allow for the subjects to return to a neutral emotional state as not to influence consecutive sessions. The stimulus videos were selected to elicit either a neutral response or one of five emotions; amusement, joy, disgust, fear and sadness.

For each of these sessions, multiple video recordings were made in the lab envi-

<table>
<thead>
<tr>
<th>Environment</th>
<th>MAHNOB-HCI</th>
<th>BINED</th>
<th>MIME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic</td>
<td>laboratory</td>
<td>majority of undergrad students</td>
<td>living room-like</td>
</tr>
<tr>
<td># of subjects</td>
<td>30 (17 female)</td>
<td>74 (42 female)</td>
<td>42 (24 female)</td>
</tr>
<tr>
<td>Avg. subject age</td>
<td>26.1 (σ = 4.4)</td>
<td>23.8 (σ = 6.9)</td>
<td>67.3 (σ = 4.2)</td>
</tr>
<tr>
<td>Videos per subject</td>
<td>20</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Stimulus type</td>
<td>pop culture</td>
<td>proven to induce specific emotions</td>
<td>historic television</td>
</tr>
<tr>
<td>Avg. video duration</td>
<td>81.4s (σ = 22.46)</td>
<td>60s (σ = 0)</td>
<td>60s (σ = 0)</td>
</tr>
</tbody>
</table>

Table 3.1: Overview of dataset characteristics

section of this chapter.

3.1 MAHNOB-HCI

The MAHNOB-HCI dataset, short for Multimodal Analysis of Human Nonverbal Behavior: Human Computer Interfaces, was presented and developed by Soleymani et al. [2012]. It is a dataset developed for the purpose of affect recognition and implicit tagging research, the latter is the concept of automatically assigning labels (tags) without manual interaction. The dataset contains recordings of 30 subjects, who each watched 20 stimulus videos in a lab environment. These subjects were of various nationalities, living in or nearby London, 17 of which are female and 13 male. The stimulus videos were obtained from popular movies as well as sites such as YouTube.com and shown in random order. Additionally neutral video clips were shown in between videos in order to allow for the subjects to return to a neutral emotional state as not to influence consecutive sessions. The stimulus videos were selected to elicit either a neutral response or one of five emotions; amusement, joy, disgust, fear and sadness.

For each of these sessions, multiple video recordings were made in the lab envi-
Datasets

3.2 BINED

Sneddon et al. [2012] developed the Belfast Induced Natural Emotion Database, which is a corpus of video recordings of subjects who performed various emotion elicitation tasks. The tasks the subjects conducted were chosen to elicit emotion in various ways, these ranged from touching an object marked to be dangerous to watching a funny video clip. For the application of emotion detection in TV viewers we selected the relevant tasks.

The resulting set from the BINED dataset contains videos from 82 subjects from North Ireland, 45 of which are female. Each watched four video clips in a semi-lab environment and provided their emotional self-assessment after each clip. The emotional self-assessment was obtained by allowing participants to provide emotional labels. In addition, a valence score was provided on a scale of -100 to 100, which we used as the actual self-assessment rating.

The stimulus videos selected for this experiment were chosen to elicit sadness (clip from a movie), disgust (clip showing surgery), anger (extract from a movie) and amusement (clip from a comedy show). Additionally neutral clips (of swimming fish, extracted from documentary) were shown to prevent emotions from carrying over between sessions.

---

1Set 2, task 5, 6, 7 & 8.
3.3 MIME

The final dataset has been collected at TNO as part of the MemoryBanks Interactive Metadata Extraction (MIME) research project. This corpus contains video recordings of Dutch elderly people who watched selected fragments from historic Dutch television programs. These subjects were selected based on their age and all lived in the Netherlands.

Using a Microsoft Kinect device, audio, video, and depth were recorded during the playback of stimulus videos. After each of twelve videos, subjects selected the most appropriate emotional label out of seven choices. A total of 42 subjects participated, of which thirteen people watched the videos solo, whereas fifteen couples watched the fragments together.

For these experiments an office room was fitted with a couch, some decoration and a regular television, resulting in a near-natural environment resembling a living room, as can be seen in Figure 3.4. After each video, the subjects were asked to pick the most appropriate emotional label out of seven options to describe their feelings during the video fragment. This was done either by selection an option with a mouse or by marking the relevant option on paper.

In order to be able to use the videos of sessions that consisting of two subjects watching at the same time, the videos were split in half. This resulted in videos showing only a single subject to avoid confusion with the facial expression recognition software which would be used later on.

3.4 Data selection

In order to ensure that the quality of each recorded session would be sufficient for our purposes, we formulated a list of requirements which were used to filter the sessions in each dataset. These exclusion criteria were as follows:

1. There is no face present in the video recording or no face could be detected.

Figure 3.3: Sample frame and photo of the experimental setup from the BINED dataset. [Sneddon et al., 2012]
Datasets

3.4 Data selection

Figure 3.4: MIME experiments.

(a) Sample frame. (b) Experimental environment.

Table 3.2: Post-exclusion dataset composition.

<table>
<thead>
<tr>
<th></th>
<th>MAHNOB-HCI</th>
<th>BINED³</th>
<th>MIME⁵</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subjects (female)</td>
<td>24 (15)</td>
<td>74 (42)</td>
<td>19 (11)</td>
</tr>
<tr>
<td>Number of sessions</td>
<td>480</td>
<td>296</td>
<td>205</td>
</tr>
<tr>
<td>Valence (low/high)</td>
<td>307 / 173</td>
<td>229 / 67</td>
<td>62 / 143</td>
</tr>
<tr>
<td>Arousal (low/high)</td>
<td>294 / 186</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Control (low/high)</td>
<td>255 / 225</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

2. No emotion self-assessment rating is available.

3. The video contains many artifacts or stutters.

Applying these exclusion criteria results in the composition as shown in Table 3.2. We excluded 8 subjects entirely from the BINED dataset because the facial expression recognition software could not detect a face in their videos. Finally, for MIME we removed 2 subjects for this very reason. Furthermore, twenty subjects were removed because no self-assessment ratings or video recordings were available for their sessions at the time we conducted the experiments reported in this work. The resulting compositions are shown in Table 3.2. A complete overview of the excluded sessions can be found in Appendix B.
Chapter 4

Automatic and Human Labeling

In the previous chapter we have seen that there exist several datasets with different characteristics. These datasets can be used to conduct emotion recognition research. In this chapter we move on to the core of our challenge; using facial expressions to determine a specific user’s experienced emotion while watching a video clip.

Various applications exist for this type of affect recognition from video recordings. One example of this would be smart TV applications, where using the viewer’s emotion, it would be possible to provide suitable video recommendations. Using facial expressions humans show their emotions, which we can register using video cameras.

In Section 4.1 we will describe the automatic methods considered in this work, which rely on the use of these datasets. While automatic methods are able to detect human emotions up to a certain point, humans are experts in the recognition of emotions in others. Being able to compare the different accuracy scores obtained by human observers with the results from our automatic methods allows us to determine how well our software manages to perform this same task. To allow us to make such a comparison, we have also conducted experiments with human annotators. The setup used to conduct this human annotation task is described in Section 4.2.

4.1 Automatic classification

In this section we start by describing the typical pipeline used in automatic emotion classification methods. This provides us with the framework to discuss the different components of methods discussed later. The first method discussed is the reference method used throughout this work, which was chosen for its use of a public dataset, allowing for comparisons between implementations. In order to maximize real-world applicability, we focused on data that can be collected without the need for sensors that need to be physically attached to the subjects. This led us to a slightly different implementation of this method, which is described next.

Since we have also created an intensity-based method, a detailed overview of the changes when compared to the previous method are outlined. Both methods however perform binary classification in to “low” and “high” classes to allow for a comparison with the earlier works and allow for application on each of three datasets shown earlier. Finally we discuss extra features that have been considered.
4.1 Automatic classification

4.1.1 Pipeline overview

As automatic emotion classification methods typically consist of an elaborate chain of different processing steps, we have identified several of the components that are used and provided a global overview of these steps. These components together make up the “pipeline” that on one end accepts the various recordings of a subject and on the other end produces a prediction of the class of this session. In general, the pipeline used by the methods in this work is shown in Figure 4.1, with each step explained below.

**Data recording** While the user watches a video, various sensor streams are recorded using video cameras and microphones.

**Data conversion** In order to convert these raw recordings into usable data, a data conversion step is applied. In this step, the raw data is converted into a more concise format. Examples of this step would be to use speech interpretation software to produce a transcript of any spoken text or using facial expression analysis software to produce a raw log of the subject’s activity during the session.
Feature extraction Once the data has been converted into a more usable format, features can be extracted from using software. For example, we could calculate the rate of change in the angle of a subject’s head during a session or construct a feature to calculate the combined activation of certain facial expressions.

Feature vector construction Using the values obtained in the previous step, aggregation functions are applied to turn these measurements over time into a limited number of scalar values, resulting in a single vector for each session, each with identical dimensions.

Feature vector optimization Before using the constructed feature vectors, we may first optimize our vectors. One such approach would be to determine which features should be included to perform classification and which features could be excluded, which may speed up and improve the classification process. In addition any scaling or normalization of the vectors may occur during this step.

Classification The resulting feature vectors can then be used to train a machine learning algorithm to recognize patterns in the subject’s behavior that corresponds to certain classes. Depending on the preferred application and presence of training data, the algorithm could either be tailored towards the behavior of one specific person or towards generic behavior. After this, the classifier can be used to predict labels for new feature vectors, using this prior information.

4.1.2 Onset/Offset Counting

As a starting point we have considered the work by Koelstra and Patras [2012]. The method they describe is based on onset/offset counting of various combinations of facial expressions, using a Gaussian Naive Bayes (GNB) machine learning algorithm at its core. We will therefore refer to this method as OOC-GNB.

They presented a method for binary classification on valence, arousal and control using different modalities, for which we used the method using facial expressions as a starting point. In their work, the authors use the MAHNOB-HCI dataset and reduced the self-assessment ratings provided in this dataset to a low class (ratings 1-5) and a high class (ratings 6-9). The implication of this reduction of a 9-point Likert scale to a binary class is that the “neutral” (5) ratings have been grouped into the “low” class.

The data conversion step in this method is handled by facial expression recognition software, presented in Koelstra et al. [2010], which is able to detect the onset and offset of 18 different Action Units. On the topic of feature extraction the authors reasoned that, since certain Action Unit combinations are associated with emotion according to the EMFACS method, these combinations would be included in their method. For each of these combinations, they determined whether the combination was active or inactive. From the facial expression data they performed feature vector construction. To do so, they then calculated the number of times a facial expression combination started (onset) or stopped (offset) as well as the mean difference between the onset and offsets.

The feature vector optimization was performed using recursive feature elimination (RFE), which selects the relevant features for classification and removes the other features. In this elimination method, a linear SVM is trained after which the 10% lowest-
In our implementation of the reference method outlined in the previous section, some changes were made. Rather than using the academic facial expression recognition software, we used the commercial Noldus FaceReader 5 software\(^1\) in the data conversion step. This software was first presented in Den Uyl and Van Kuilenburg [2005] and allowed us to monitor two additional Action Units in comparison to the method used in the original OOC-GNB method. Additionally, rather than providing us with onset and offset events, the FaceReader software reports the intensity level of each Action Unit, which can either be “inactive” or range from A to E, with E being most intense. Using the intensity level we determined the onset and offset events in terms of the intensity increasing or decreasing during measurement intervals, effectively reproducing the data reported by the software in the original OOC-GNB method.

Furthermore, in the original work a GNB classifier was used in the classification step, the probabilistic output of which was later used to fuse multiple classifiers. In our preliminary research we found that our implementation of the OOC-GNB method failed to meet the performance presented in the work by Koelstra et al. [2010] (Appendix C.1). Since we focused primarily on the usability of facial expressions, the need for the probabilistic output was no longer present, allowing us to use a different machine learning algorithm. Based on the results published by Senechal et al. [2014] and the popularity of SVM in related work, we used a linear SVM for both feature

---

\(^1\)[http://www.noldus.com/human-behavior-research/products/facereader]
selection and classification. We will refer to this implementation as OOC-SVM. The implementation of this SVM was provided by the open source Python library scikit-learn described in Pedregosa et al. [2011].

The feature vector optimization method used by Koelstra and Patras consisted solely of applying the RFE algorithm, where the number of features chosen was found using a 10-fold cross-validation on a training set. This training set was presumably the MMI dataset [Pantic et al., 2005] used to train their facial expression recognition software. As we did not have access to this dataset and required no additional training for our facial expression recognition software, we instead used a training set of 6 randomly selected subjects, which were excluded from further experiments, to find this number. Furthermore, we evaluated the use of feature vector scaling, as used in many works we discussed in our related work chapter. We used the MinMaxScaler provided by the scikit-learn library for this task, which applies the following formula for each feature value $x$ to scale it to $[0, 1]$:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Finally for the classification step the training method used in the reference implementation was altered. In the original work, only personalized (leave-one-session-out) cross-validation was evaluated in which the system is trained specifically using the history of one subject. We reasoned that these methods can only work in a real-life setting if the system does not require training for one specific person. In order to evaluate the performance of this change, we also used global training (leave-one-subject-out) cross-validation. This results in the machine learning algorithm being trained on, for instance 23 subjects, followed by the prediction of all ratings for the 24th subject.

### 4.1.4 Facial expression intensity method

With the reference method outlined in the previous sections as a starting point, we created a new method. This method is based on the intensity with which a facial expression is shown, prompting the name Facial Expression Intensity (FEI). Recently the use of these expression intensities was also mentioned by Senechal et al. [2014] as direction for future work. The goal of the adjustments made was to improve the classification performance, improving the usability of these techniques for real-life applications. First we present our intensity-based method, which uses different feature extraction and feature vector construction techniques.

For our intensity-based method we first changed the feature extraction step to consider the intensity of each Action Unit reported by the expression recognition software. To do so we converted the reported intensities to integer values (0 to 5). Our second adaptation was to change the feature vector construction step. Rather than counting the number of occurrences, we calculate the average and the sample standard deviation. This resulted in a feature vector consisting of $22 \times 2 = 44$ features. An advantage of this approach is that by not relying on counting, the method is invariant for the length of a session.

Similarly to our OOC implementation outlined in the previous section, we evaluated this method using two machine learning techniques. First we evaluated the use of personalized training to allow for a comparison with the two OOC methods, after which we also evaluated the use of global training.
4.1.5 Additional features

In addition to video footage, the MAHNOB-HCI dataset also contains ECG measurements. This meant that accessing the participant’s heart rate was also possible in our evaluation. Since solutions to extract the heart rate of a person using only video recordings are available\(^2\), the use of intrusive sensors placed on the subject’s body might no longer be necessary for heart rate detection in the future. We therefore decided to experiment with the inclusion of heart rate and heart rate variance features in anticipation of more flexible video-based solutions.

We used the EDFbrowser software\(^3\) to provide us with the time between each heart beat, as measured by the difference between highest peaks (RR-interval). From these RR-intervals, we then calculated the heart rate which in turn was used to find the average and sample standard deviation of the heart rate. Additionally we calculated the heart rate variability using the root mean square of successive differences (RMSSD). This method has been used in the context of emotion recognition [Lee et al., 2005] and its formula is as follows:

\[
RMSSD = \sqrt{\frac{\sum_{i=1}^{N} (RR_i - RR_{i+1})^2}{N-1}}
\]

When using video-based heart rate detection these methods can likely be applied as well, since RR-intervals and beats per minute (obtained from video analysis) are convertible using the formula \(HR = 60/RR\).

Finally, features regarding the head orientation could also be included, as the FaceReader software provides estimates of the subject’s orientation using video analysis. We reasoned that if a person is not interested in viewing a particular video clip, it is likely that their gaze wanders, resulting in a change in the head orientation. To facilitate this we produced six new features: the average and standard deviation of the X, Y and Z-axis of the subject’s head orientation.

4.2 Human classification\(^4\)

In addition to the automatic approaches, we have also conducted experiments to determine the accuracy of human annotators when observing the sessions from the three datasets. The primary goal of this experiment has been to determine the performance of human annotators on the given datasets. We reason that humans have many years of experience in assessing the emotions of their peers, measuring their performance in comparison to automatic methods might a clue as to what the practical limitations of the use of facial expressions might be. In the following sections we will outline the software used as well as the demographics of our participants.

---

\(^2\) [https://github.com/thearn/webcam-pulse-detector](https://github.com/thearn/webcam-pulse-detector)

\(^3\) [http://www.teuniz.net/edfbrowser/](http://www.teuniz.net/edfbrowser/)

\(^4\) An extended abstract of this work was accepted for the Measuring Behavior 2014 conference [Holkamp and Schavemaker, 2014].
4.2 Human classification

A web-based video annotation tool was developed to allow us to easily recruit participants and allowing the participants to perform the annotation task from their own computer. The process was as follows; when a participant visited the annotation task using their browser, we presented them with instructions regarding the experiment. The participant could then click “Start”, after which they were asked to provide their age group, gender and nationality. Once this data was provided, the participants could continue on to the actual annotation task. At this point, a video was shown, with on the right hand side of the video two questions, it was possible for the subject to play the video multiple times. The first question asked the participant to provide an estimation of the emotion of the person shown in the video, using a 5-point Likert-scale, ranging from “negative” to “positive”. We explicitly instructed the subjects to choose “neutral” if they could not distinguish an emotion. Furthermore we asked to rate how visible the subject’s emotion was in their facial expression, using a 5-point Likert-scale ranging from “not visible” to “very explicit”, allowing for future analysis. The interface of the annotation task is shown in Figure 4.2.

To increase the number of sessions that could be annotated, videos were played at four times the normal speed, which was reported by Kamachi et al. [2001], Bould et al. [2008] to have limited impact on detection accuracy. Moreover, no audio was included in the videos to provide our human annotators with the same information used by the automatic classifier. In order to prevent exhaustion, we presented the raters with a page that encouraged taking a break at 4-5 minute intervals during the experiment, however, this was not enforced.

4.2.2 Annotation method

In three separate experiments, one per dataset mentioned in Chapter 3, we asked the participants to annotate a sample from each dataset. The number of videos was cho-
Table 4.2: Human annotation participant demographics: number of participants for each category is shown.

To determine the performance of our annotators, we reduced their input to binary, similar to what has been done with the self-assessment ratings for the three datasets. This results in a class "low" (ratings 1-3) and "high" (ratings 4-5). We then compared the majority vote per video with the ground truth as provided through the self-assessment ratings. In addition, we determined the average annotator performance by taking the average of the overall accuracy of each annotator.

Participants

The annotation task was spread among employees of the Media and Network Services department at TNO, students and staff of the Web Information Systems department of Delft University of Technology as well as friends via social media. Although the participants were free to stop at any time during the annotation task, the majority fully annotated at least one of the datasets in its entirety. Due to this, we decided to only include the annotations from those participants who fully completed the annotation task, mitigating the effect of users who provided only a small number of ratings or only completed the training videos.

For the first dataset, we have found 15 participants willing to complete our full annotation task, who were all 21-40 years old and mainly of Dutch nationality, in addition to one Swedish and one French participant. Similar numbers were obtained for the second dataset, for which 14 participants completed the annotation task, and the third dataset, with 10 participants. A complete overview of the demographics of our participants is shown in Table 4.2.

For the first experiment we selected two fragments for random stimuli for each subject in the MAHNOB-HCI dataset, resulting in a total of 48 videos.

The second experiment contained a sample of the BINED dataset. This sample consisted of a stratified random sample of 36 sessions as a sample of two fragments per subject was unattainable given the large number of subjects. In our sample, 20 sessions contained female and 16 contained male subjects, roughly corresponding to the ratio present in the original dataset.

The third and last experiment consisted of videos from the MIME dataset, from which we randomly sampled 25 sessions (either solo or duos). For the trials where two subjects were present, we split the video recording into two parts showing a single person, which we both included. This resulted in a total of 32 videos, of which 16 videos consisted of both subjects of a single pair and 16 videos of solo subjects.
Chapter 5

Evaluation

Now that we have seen several methods of both automatic and human annotation of emotion, we present the evaluation of our results. First, in Section 5.1, we describe the evaluation of our automatic experiments, including the metrics that have been used, the experimental setup and the results we obtained during each experiment.

Next, in Section 5.2, the human classification experiments are described and the obtained results for each of the three datasets are presented. The results of both the software and human annotation experiments are then shown in Section 5.3, providing a “man vs machine” comparison.

5.1 Automatic classification

In this section we describe the results and experiments obtained with our machine-learning classification methods. In our analysis, we have included two general baseline methods. The first method was a random classifier, which produces one of two classes at random. The second method was a classifier returning the majority class, which always returns the most common class for a given dataset. Once we had these baseline methods set up, we compared the performance against the implementations we presented in Section 4.1. In our comparisons we have used two metrics, the classification rate and average F1-score.

First we compared the performance figures reported of the onset/offset counting (OOC) method. Second the results of our facial expression intensity (FEI) method are presented and compared to the previous results. Next we have evaluated the suitability of both of these methods when training for a global group of subjects, rather than one specific subject. Finally the results of the use of additional features are discussed, for which we considered both head orientation and heart rate either separately or each combined with the FEI features.

5.1.1 Metrics

To evaluate the performance of the methods discussed in this chapter, we have used two primary metrics. These metrics are the F1-score and classification rate or accuracy.

The accuracy of the method is calculated by comparing the predicted class for each session with the ground truth value provided by the subject in the dataset. The
inclusion of accuracy allows us for comparisons with other research as well as the human annotator performance presented later in this chapter. The accuracy score is however affected by unbalanced datasets, where one class might be a lot more common than the other.

To avoid this factor and allow for comparisons with other works, we have also included the F1-score. The formula used to calculate this score, shown below, is equal to the weighted average of the precision and recall and can also be expressed in terms of true positives, false negatives and false positives.

\[
F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \cdot \text{true positive}}{2 \cdot \text{true positive} + \text{false negative} + \text{false positive}}
\]

In particular, we have used the average F1-score, which is equal to the average of the F1 score for each class. The formula used is as follows, where \( N \) is equal to the total number of entries and \( F_{1\text{low}} \) is the F1-score calculated with the “low” class as correct or positive class:

\[
F_{\text{average}} = \frac{F_{1\text{low}} + F_{1\text{high}}}{2}
\]

Using this metric we accounted for the fact that the datasets used have unbalanced classes, i.e. more “low” than “high” sessions. By taking the unweighted average of the two classes we ensure that a class unbalance in the dataset will not lead to our algorithms obtaining a high F1-score while only being able to correctly classify the most common cases. Alternatively, an inversely weighted average F1-score could be considered, which takes the average of the F1-scores for each class, inversely weighted to the size of the class. This further emphasizes the classification performance of the smallest class. To allow for comparisons between our work and Koelstra and Patras [2012] we have however used the average F1-score.

### 5.1.2 Onset/offset counting implementations

**Setup**

In our first experiment, we have evaluated the performance of the onset/offset counting (OOC) method. This method was presented by Koelstra and Patras [2012] and relies on the tracking of various Action Unit combinations known to relate to emotion. The detection of facial activity was performed using the Noldus FaceReader software, version 5.1. The software was configured to use the “general” training preset. The number of times these combinations were observed to start and stop are then included into a feature vector. The machine learning algorithm is then trained using the personalized method, which leave-one-session-out cross-validation for a specific subject. We have implemented this method as well, with small adjustments, as is described in Section 4.1.3.

This comparison helps us identify whether our implementation, OOC-SVM, obtained similar performance to the reference method. We have evaluated this method using the MAHNOB-HCI dataset, outlined in Section 3.1, and classification was performed for the valence, arousal and control targets of each session. In addition to the
### Evaluation

#### 5.1 Automatic classification

<table>
<thead>
<tr>
<th>Method</th>
<th>Valence F1</th>
<th>Valence CR</th>
<th>Arousal F1</th>
<th>Arousal CR</th>
<th>Control F1</th>
<th>Control CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>OOC-GNB - Koelstra and Patras</td>
<td>63.3 64.0</td>
<td><strong>66.3</strong> 67.5</td>
<td>62.0 62.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OOC-SVM</td>
<td>64.7 66.2</td>
<td>58.1 61.3</td>
<td>61.3 61.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEI</td>
<td><strong>67.6</strong> 69.8</td>
<td>60.7 62.7</td>
<td><strong>64.8</strong> 65.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random</td>
<td>48.7 50.0</td>
<td>48.1 50.0</td>
<td>48.2 50.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Majority class</td>
<td>38.3 62.6</td>
<td>37.6 62.0</td>
<td>39.4 53.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: Classification rates (CR) and average F1-scores for both implementations of the OOC method. The methods were applied on the MAHNOB-HCI dataset. The highest F1-score per classification target is shown in bold.

reference implementation of OOC and OOC-SVM, we have also included the performance of a randomized and majority class classifier.

**Results**

The full collection of results is shown in Table 5.1. We observe that the OOC-SVM method obtains a slight improvement over the OOC-GNB performance for the valence class. The performance for the control label is however lower, at 61.3 for OOC-SVM versus 62.0 for OOC-GNB, a difference of less than 2%. The classification score for the arousal label is however 9% lower, for both the average F1-score and classification rate.

#### 5.1.3 Facial expression intensity

**Setup**

Next we evaluated the performance of our facial expression intensity-based method (FEI). In contrast to the OOC methods described earlier, this method relies on the intensity of each detected Action Unit. From these intensities, the average and standard deviation are calculated and included in a feature vector. The machine learning algorithm is then trained in the same way as with the OOC methods. A more detailed description of this method is provided in Section 4.1.4.

This method was evaluated in a similar manner to the earlier methods. Classification was performed for the valence, arousal and control labels present in the MAHNOB-HCI dataset.

**Results**

We compared these results with the results found for the OOC methods in the previous section. The results are shown in Table 5.1. We can observe that for all three targets, the F1-scores obtained by the FEI method surpass our OOC-SVM method. In addition to this, for both the valence and control targets, the F1-score obtained by the FEI method surpasses the performance of the OOC-GNB implementation by Koelstra and Patras.
5.1 Automatic classification

### Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Personalized training</th>
<th>Global training</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>CR</td>
</tr>
<tr>
<td>OOC-SVM</td>
<td>64.7</td>
<td>66.2</td>
</tr>
<tr>
<td>FEI</td>
<td>67.6</td>
<td>69.8</td>
</tr>
</tbody>
</table>

Table 5.2: Comparison of F1-scores for valence classification using different training methods on the MAHNOB-HCI dataset.

#### 5.1.4 Training methods

A comparison between personalized and global training has also been made, the difference being whether the machine learning algorithm is trained to specifically recognize the performance of one person versus recognizing the general behavior of all subjects. A more detailed description of the differences between these training methods is outlined in Section 4.1.4.

#### Setup

For the first experiment we set up both the OOC-SVM and FEI methods to first train using the personalized method used earlier. We then changed the pipeline to train in a global fashion. This meant that the system was first trained on the sessions of all subjects except for one subject, after which the system was evaluated using the sessions of this subject through cross-validation. We have used the MAHNOB-HCI dataset where classification was performed on the valence class.

After we have conducted these experiments, we have also evaluated the performance of the OOC-SVM and FEI methods on the BINED and MIME datasets. These datasets contained insufficient sessions per subject to allow for personalized training. As only the MAHNOB-HCI dataset has arousal and control ground truth values available, for this experiment classification was performed for only the valence label. The FaceReader software was configured to use the training data for elderly people when processing the videos of the MIME dataset and set to “general” for the other datasets.

#### Results

The results of the first experiment are shown in Table 5.2. Here we can see that the FEI method obtains a lower F1-score when using global training on the MAHNOB-HCI dataset rather than personalized training. On the other hand, the performance of our OOC-SVM implementation appeared to be more consistent. In addition, where the classification rate of the FEI method was decreased by almost two percent points, the OOC-SVM method saw an increased by four percent points.

The results of the second comparison made are shown in Table 5.3. We observe that the F1-scores obtained by the FEI method are higher than the scores achieved by the other methods. This behavior is however not reflected in classification rate for both the MAHNOB-HCI and MIME dataset, where the highest rates are instead obtained by the OOC-SVM and majority class baseline, respectively. We observe furthermore that for the MIME dataset, the classification rate obtained by the FEI method is lower than the majority class performance of this dataset.
Table 5.3: Classification rates and F1-scores for binary classification of valence using global training. The highest F1-scores and classification rates per dataset are shown in bold.

5.1.5 Feature vector optimization

Next we evaluated the impact of different feature vector optimization methods. For this we have conducted two experiments, which are described in this section.

Setup

In our first experiment, we evaluated the impact of using recursive feature elimination (RFE) on the classification performance of the OOC-SVM and FEI methods outlined earlier. We first trained the two methods using personalized training. Here both methods were provided with all but one session of a specific subject from the MAHNOB-HCI dataset. In this case binary classification was performed for valence, arousal and control targets and the methods were executed once with and once without RFE enabled in our pipeline. As a second step we evaluated the performance of both methods using global training, where the methods were trained using the behavior of all but one users of a specific dataset and tested on the single user. We again evaluated the binary classification performance, both with and without RFE enabled. In this case only the valence target was used, due to the absence of the other two labels in the BINED and MIME datasets.

Our second experiment aimed at evaluating the impact of feature vector scaling on the classification performance. We used an approach similar to the first experiment, in this case we compared the performance both with and without feature scaling. When scaling was applied, the feature values were scaled to the [0, 1] range, as described in Section 4.1.3.

Results

The results of the first comparison can be found in Table 5.4. In Table 5.4a we see that the use of RFE proves to be beneficial, resulting in small increases in F1-scores when compared to the methods without RFE enabled. There was is one exception visible, which is the arousal classification using the OOC-SVM method. This combination obtained an F1-score of around 1.5% lower than the same combination with RFE enabled. In the other combinations, the methods with RFE enabled obtained higher or equal scores to the methods without RFE enabled. The largest delta being 10% increase in the F1-score for control classification.
### 5.1 Automatic classification

#### Table 5.4: Comparison of classification performance with and without RFE. The highest F1-score per target is shown in bold for each method pair.

<table>
<thead>
<tr>
<th>Method</th>
<th>RFE</th>
<th>Valence</th>
<th>Arousal</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OOC-SVM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td>64.7</td>
<td>66.2</td>
<td>58.1</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>64.7</td>
<td>66.2</td>
<td><strong>58.9</strong></td>
</tr>
<tr>
<td><strong>FEI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td><strong>67.6</strong></td>
<td>69.8</td>
<td><strong>60.7</strong></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>67.5</td>
<td>69.6</td>
<td>59.2</td>
</tr>
</tbody>
</table>

(a) Personalized training

<table>
<thead>
<tr>
<th>Method</th>
<th>Scaling</th>
<th>Valence</th>
<th>Arousal</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OOC-SVM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td>61.2</td>
<td>68.8</td>
<td>70.9</td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>61.2</td>
<td>68.8</td>
<td><strong>71.7</strong></td>
</tr>
<tr>
<td><strong>FEI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td><strong>64.8</strong></td>
<td>66.7</td>
<td><strong>74.2</strong></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>64.1</td>
<td>66.0</td>
<td><strong>74.2</strong></td>
</tr>
</tbody>
</table>

(b) Global training

#### Table 5.5: Comparison of classification performance with and without feature vector scaling. The highest F1-score per target is shown in bold for each method pair.

<table>
<thead>
<tr>
<th>Method</th>
<th>Scaling</th>
<th>Valence</th>
<th>Arousal</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OOC-SVM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td>64.7</td>
<td>66.2</td>
<td><strong>60.5</strong></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>61.7</td>
<td>64.0</td>
<td>53.7</td>
</tr>
<tr>
<td><strong>FEI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td><strong>67.6</strong></td>
<td>69.8</td>
<td><strong>60.7</strong></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>65.8</td>
<td>67.9</td>
<td>54.8</td>
</tr>
</tbody>
</table>

(a) Personalized training

<table>
<thead>
<tr>
<th>Method</th>
<th>Scaling</th>
<th>MAHNOB-HCI</th>
<th>BINED</th>
<th>MIME</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OOC-SVM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td>61.2</td>
<td>68.8</td>
<td><strong>70.9</strong></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>59.7</td>
<td>64.2</td>
<td>67.1</td>
</tr>
<tr>
<td><strong>FEI</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes</td>
<td></td>
<td><strong>64.8</strong></td>
<td>66.7</td>
<td><strong>74.2</strong></td>
</tr>
<tr>
<td>no</td>
<td></td>
<td>63.9</td>
<td>65.6</td>
<td>73.1</td>
</tr>
</tbody>
</table>

(b) Global training

In Table 5.4b, we see that the results obtained when performing global training are more diverse. The results remain equal whether or not RFE was performed for the OOC-SVM method on the MAHNOB-HCI dataset and the FEI method on the BINED dataset. In three of the four remaining cases, the absence of RFE resulted in an improvement in classifier performance.

The second experiment, comparing the use of feature vector scaling, resulted in the classification scores shown in Table 5.5. We can see that, with two exceptions,
5.1 Automatic classification

<table>
<thead>
<tr>
<th>Method</th>
<th>MAHNOB-HCI</th>
<th>BINED</th>
<th>MIME</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>CR</td>
<td>F1</td>
</tr>
<tr>
<td>Head Orientation (HO)</td>
<td>44.6</td>
<td>44.6</td>
<td>57.7</td>
</tr>
<tr>
<td>FEI</td>
<td><strong>64.8</strong></td>
<td>66.7</td>
<td>74.2</td>
</tr>
<tr>
<td>FEI+HO</td>
<td>63.4</td>
<td>64.8</td>
<td><strong>76.4</strong></td>
</tr>
</tbody>
</table>

Table 5.6: Global training: Classification rates and F1-scores for various combinations of head orientation and intensity-based features. The highest F1-scores per dataset are shown in bold.

the use of scaling improved the F1-scores of the methods. The first exception is the performance of the OOC-SVM method for the control label (Table 5.5a), here the F1-score was reduced from 60.6 to 56.3 in the RFE-enabled case. The second exception is the FEI method on the MIME dataset (Table 5.5b), where the F1-score was reduced from 51.7 to 51.0 when RFE was used, a difference of less than two percent.

5.1.6 Additional features

Extra features have also been considered and evaluated in our experiments. These features were either based on the head orientation of the subject, which we will discuss first, or heart rate features based on available ECG data. Detailed descriptions of both methods are provided in Section 4.1.5.

Setup

The first experiment conducted was the use of a set of features based on the head orientation of a subject. These data were provided by the Noldus FaceReader software, which provided the orientation in number of degrees in the X, Y and Z axis. These were converted into an average as well as standard deviation of each axis and used in the same pipeline as used for the FEI method. We have evaluated both the sole use of these features as well as in feature-level fusion with FEI method features. This method was evaluated using all three datasets, for which classification was performed for the valence target after performing global training.

The second experiment conducted was the inclusion of various heart rate (HR) features, such as the average and standard deviation of beats per minute. It has been evaluated with both personalized training as well as global training. In both cases the MAHNOB-HCI dataset was used where classification was performed for valence, arousal and control target labels. It has been evaluated using both personalized and global training methods.

Results

For the first experiment, the results of the inclusion of the head orientation (HO) features are shown in Table 5.6. We observe that the inclusion of the head orientation feature did not result in higher performance for the MAHNOB-HCI dataset. On the other two datasets, we however observed an improvement of the head orientation fusion with FEI over the FEI method.
### 5.2 Human annotations

In our human annotation experiments, we gathered a crowd of volunteers to perform annotation tasks. In these tasks, the annotators were asked to assess the valence (positive or negative emotion) experienced by the subjects of each of the three datasets. These tasks were completed by our annotators from their own computers, without compensation for their effort. A detailed description of the method used as well as the demographics of our crowd can be found in Section 4.2.

#### Setup

We reduced the input provided by our annotators to two classes, “low” and “high” to allow for a comparison with the ground truth used in the machine learning classification. The method to achieve this was similar to the one used to process the self-assessment ratings for the MAHNOB-HCI dataset. This results in a class “low” (ratings 1-3) and “high” (ratings 4-5). We then compared both the majority vote per video with the ground truth as provided through the self-assessment ratings. The number of correct ratings divided by the total number of ratings then provides us with the overall accuracy of the annotator. In addition, we determined the average annotator performance by taking the average of the overall accuracy of each annotator.

Furthermore we have determined the inter-rater agreement per dataset, as this might provide an indication of the difficulty of the assessment. The metric we used to calculate this is Fleiss’ kappa, as described by Landis et al. [1977].

#### Table 5.7: Comparison of valence classification: the effect of different training methods and feature combinations on the F1-scores. The highest scores per training method are shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>Personalized</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR-features</td>
<td>49.0</td>
<td>45.2</td>
</tr>
<tr>
<td>FEI</td>
<td>67.6</td>
<td>64.8</td>
</tr>
<tr>
<td>HR+FEI</td>
<td><strong>68.1</strong></td>
<td>62.0</td>
</tr>
</tbody>
</table>

For the second experiment we only include the results for the valence feature in Table 5.7 for both personalized and global training for brevity. The full set of results can however be found in Appendix C, Table C.2a & b.

When we consider the results shown in Table 5.7, we observe that for the personalized training technique, the inclusion of HR features improved the F1-scores obtained for valence classification, resulting in an F1-score of 68.1 versus 67.6 for only the FEI method. This behavior was not observed for the arousal and control dimension. When using global training, this improvement was not found for the valence and arousal targets. An increase in F1-score from 51.0 to 56.6 was observed however for the feature-level fusion of the HR and FEI features on the control label.
Results

An overview of the findings of these experiments can be seen in Figure 5.1, where the average classification rates of the human annotators are shown next to the expected performance of a random and majority class recommender. What we observe here is that for both the MAHNOB-HCI and BINED datasets, the human annotators obtain the highest classification rate, followed by the FEI method. The exception in this overview is the MIME dataset, where the majority class classification rate surpasses the other two methods. We found no significant differences in performance between the different age groups, genders or nationalities among our raters.

In order to evaluate the inter-rater agreement for each dataset, we have calculated the Fleiss’ kappa. The results of which are shown in Table 5.8, a possible interpretation, as described by Landis et al. [1977], is also included. We can observe that overall, our annotators obtained at least a moderate agreement. Furthermore, we can observe that the agreement on the BINED dataset was highest, at a kappa of 0.88. On the other hand, the agreement regarding ratings on the MIME dataset was lower at 0.52.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Fleiss’ kappa</th>
<th>Landis et al. interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAHNOB-HCI</td>
<td>0.66</td>
<td>Substantial agreement</td>
</tr>
<tr>
<td>BINED</td>
<td>0.88</td>
<td>Almost perfect agreement</td>
</tr>
<tr>
<td>MIME</td>
<td>0.52</td>
<td>Moderate agreement</td>
</tr>
</tbody>
</table>

Table 5.8: Inter-rater agreement between human annotators on each of the three datasets.
5.3 Comparison of human and automatic annotation

Evaluation

Setup

Finally, we have compared the performance of our human annotators and machine learning classifiers after their individual evaluations. In particular we compare the human annotations with our intensity-based method outlined in Section 4.1.4, trained using the global training method (leave-one-subject-out). In this comparison we will consider the average accuracy of our human annotators from each experiment as well as the overall results obtained by our FEI method on the same dataset. We compare the difference in accuracy between human and machine using a two-sided pairwise T-test at $p=0.05$ confidence.

Results

Now that both the machine learning classification methods and human annotator performances have been considered, we present a comparison of these two methods. A visual representation of this comparison is shown in Figure 5.2. We observe that for both MAHNOB-HCI and BINED the human classification rate is consistently higher than the performance of both the majority class and intensity-based method. For the MIME dataset this is not the case, here the majority class recommender outperforms both the human annotators as well as the FEI method.

Using a two-sided pairwise T-test we have found no significant difference between the accuracy of our FEI method and the human annotator accuracy on each of the datasets.

Figure 5.2: Classification rates for human annotators and the intensity-based method on the three datasets considered.
Chapter 6

Discussion

In the previous chapter we presented the results of our experiments, which we will discuss in this chapter. First, we will provide a short summary of our findings.

In our first experiments, we found that our OOC-SVM method improved on the performance of the OOC-GNB by Koelstra and Patras [2012] for the valence target. Moreover, the FEI method was shown to obtain higher scores in general than the OOC-SVM method.

We evaluated the use of global training of our machine learning classifier. With this method, the system is not trained to recognize the behavior specific to one person but rather of all subjects. These changes were shown to have a small impact on classification performance, the difference between the FEI and OOC-SVM methods however remains stable.

Next we have seen that the use of feature vector optimization has an impact on the classification performance. We saw that the use of RFE proved to be beneficial in particular when using personalized training, whereas the effects were neutral at best when using global training. Furthermore we saw that the use of scaling on the features in the vector improved the classification performance, with a small number of exceptions.

In our experiments we found that the use of head orientation features did not improve the classification performance in comparison to the FEI method. The fusion of heart rate features with FEI improved the valence classification performance when using personalized training and control performance when using global training. Conversely, in the other situations no positive effects were found.

Finally the human classification experiments are considered. We found that the classification rates by our human annotators are higher than those obtained by our machine learning classifiers. The exception being the MIME dataset, where both man and machine do not improve over the majority class baseline. Inter-rater agreement found during this experiment ranged from moderate on the MIME dataset to almost perfect for BINED.

Now that we have seen a short summary of the findings, we will discuss the results we found. First we will look at the results of the automatic labeling methods, after which we will continue our discussion with the human classification as well as the comparison between human and automatic annotations.
6.1 Automatic labeling

6.1.1 Automatic labeling methods

The first results we have seen in our evaluation are the performance figures reported for our implementation of the OOC-SVM method and our FEI method (Table 5.1). Both our implementations resulted in higher F1 scores and classification rates than the figures reported in the original work by Koelstra and Patras [2012] for the valence labels. The performance figures of the two algorithms on the other two target labels did however not meet the performance reported in the original work.

This discrepancy might be caused by the choice of facial expression analysis software. It is possible that the Noldus FaceReader software we used in our experiments did not pick up on some of the facial activity shown by the subjects. This could also explain the discrepancy between the performance of the original work and the performance we obtained with our OOC-GNB implementation in our preliminary research. According to the information from the vendor, the FaceReader software is trained using video material of varying resolutions with subjects covering a wide range of ages and ethnicities. Due to this, the software may be more tailored towards the recognition of facial expressions in a wide range of subjects whereas solutions found in literature might be better able to recognize subtle expressions in a more limited range of demographics.

To the best of our knowledge, no comparisons of the FaceReader software with similar academic applications have been published. This means that at this time, we are not able to determine whether the difference in performance may be caused by does not allow us to determine whether some Action Units could explain the difference we found, at this time. We do however know that the academic facial expression recognition software used in the work by Koelstra and Patras was trained on a single dataset, whereas the FaceReader software was trained on a wider range of people, including infants and elderly people.

Alternatively, our method of determining the number of features to be selected using the RFE algorithm could limit the performance of our software as we did not calculate this number based on the performance on a different dataset as Koelstra and Patras did in their work. We found that the choice of a different number of features did however not impact the performance of our method significantly.

Positively, the results also showed that our FEI method consistently improved over both the figures for the original OOC-GNB method as well as our implementation when disregarding the arousal dimension. This suggests that the use of the facial expression intensity, as described in Joho et al. [2009], appears to have improved classification performance for these datasets and our application.

6.1.2 Training methods

After our experiments with the methods, we also evaluated the use of a different training method. When we consider the performance figures reported in experiment (Table 5.2), we see similar results as with the personalized training experiments. Although the FEI method appears to be better suited for personalized training (leave one session out) than global training (leave one subject out), it still obtains higher F1-scores
than the OOC-SVM method on the MAHNOB-HCI dataset. Interestingly, the classification rate of the OOC method does exceed the F1-score of the FEI method in the global training scenario. This suggests that while the OOC-SVM method managed to correctly classify more items, the FEI method did so more evenly distributed between the two classes, as is shown by the higher F1-score.

Using the global training method, we also evaluated the performance of both our FEI method and our OOC implementation. The results of this experiment (Table 5.3) further followed the pattern shown in the previous experiments. For each of the three different datasets considered, the FEI method again obtained F1-scores that were higher than the OOC-SVM implementation and baseline methods. Again, in some cases the classification rate of the OOC-SVM outperformed the classification rate of the FEI method. This difference between the two metrics is explained by the fact that the datasets are unbalanced in terms of the number of examples per valence label. MAHNOB-HCI and BINED contain more negative examples whereas the MIME dataset contains more positive examples.

### 6.1.3 Feature vector optimization

After we evaluated different training methods, we also evaluated the impact of several feature vector optimization techniques on the classification performance. In our evaluation we have seen that RFE seems to have a mostly positive effect on classification performance when using personalized training. For global training, we have seen that the effects were more negative. This might suggest that the relevance of specific AUs might be rather person-specific, considering that the algorithm is better able to pick up on the relevant features using the history of one specific user than on using the data from a larger number of other people.

With feature scaling to the $[0,1]$ range we found that this improved the classification performance in all but two cases, resulting in an overall positive effect. An additional advantage of the scaling step was found in the run-time of the SVM algorithm, the time required to perform both training and classification was significantly reduced after scaling the vectors.

Overall we can conclude that the “best” optimization methods to use depend for a large amount on the dataset and training method. When picking the most suitable processing steps, one should be careful to avoid picking a solution that might be over-fitted for one dataset, possibly reducing the performance of the system on other datasets.

### 6.1.4 Head orientation features

In our research we also evaluation of the use of features based on the head orientation of our subjects, which was reported by our facial expression analysis software. While we originally reasoned that a bored subject might look away from the monitor, we did not obtain positive results when evaluating the use of this feature on the MAHNOB-HCI dataset. Both when using only the head orientation features or fusing them with the features from the FEI method, the classification performance remained below that of the FEI method. We can see however that this feature improved the classification performance for both the BINED and MIME datasets (Table 5.6).
6.1 Automatic labeling

The effect might be explained by the differences in experimental setup for each of the datasets. In the MAHNOB-HCI dataset, relatively little appeared to be visible in the field of view of the subjects. Even when the subjects might not have been interested by the stimuli, they might simply not have found anything of more interest to look at, resulting in little head movement. An alternative explanation might be in the duration of the videos, which lasted for only up to two minutes, which might be too short to properly induce boredom. Furthermore the subjects of this dataset wore an EEG cap and were instructed to keep a healthy posture during the experiment for health reasons, which might have further limited their movement.

On the other hand, the BINED dataset used rather extreme stimuli, which in several cases caused subjects to look away from the monitor. In these cases, their head also moved which would be picked up by the head orientation features. For the MIME dataset, the number of people present in the room was another major difference with the other two datasets. In this dataset, some of the subjects watched the videos with another person sitting next to them. Certain videos, possible in particular those considered interesting or funny, might elicit conversations, which in turn might result in head movement to look at the other person. Furthermore, the subjects in the MIME experiments also had a mouse (first subject) or pen and paper within reach (second subject) to register their emotional ratings. After studying the video recordings we found that several subjects looked at their ratings sheet at some point during the experiment, while the stimulus video was still being shown.

6.1.5 Heart rate monitoring

In addition to the head orientation feature, we evaluated the use of features derived from heart rate measurements available in the MAHNOB-HCI dataset. In our experiments, we found an advantage to the use of HR features when using personalized training for valence classification (Table C.2). No effect was observed for the arousal and control labels.

In the case of global training, we only found a positive effect of the inclusion of HR-features for control classification, where an improvement of 10% on the F1-score was observed. In other situations, the fusion method did not produce improved results. This limited effect on arousal classification came unexpected, as Janssen et al. [2013] also reported that the HR-based features were effective in the prediction of emotion, which also included emotions with similar valence but different arousal levels. A possible explanation of this effect could be that the aggregation methods we used for our HR features did not properly capture distinctive information present in the sensor data. Proper classification using the heart rate might thus require more in-depth sensor data processing, possibly combined with noise filtering.

An alternative explanation could also be given by Brosschot and Thayer [2003]. They found that while arousal could strongly predict the initial HR, after several minutes the heart rate was solely predicted by emotional valence. In the case of our experiments this might mean that the interval during which the subject’s heart rate was monitored was too long for arousal to have a clear correlation with the observed heart rate features.

On the topic of heart rate monitoring we can conclude that the use of several features derived from the heart rate of a subject, measured using high quality apparatus,
provided a small advantage when performing valence classification, specific to one person. In other situations, the effect was not evident in our experiments, with the exception of global training for control classification. We might assume that video-based heart rate analysis, as currently under development by several large companies, will not be able to obtain measurements of similar quality to a professional ECG apparatus. This gives rise to the question whether the use of video-based heart rate will be of use in affective applications in the future.

6.2 Human labeling

In our human labeling experiment we evaluated the accuracy of a crowd of human volunteers on a subset of sessions from three different datasets. One of the interesting observations here were the inter-rater agreement scores obtained for each of the three datasets. We found that our annotators agreed very well on the proper labels on the BINED dataset, which contains subjects who watched the most extreme emotion elicitation videos of the datasets considered. This “almost perfect” agreement also shows in the classification rate of around 90% obtained by our annotators on this dataset.

More difficult datasets were the MAHNOB-HCI and MIME datasets, which both showed more subtle or “TV-like” videos. This difference also showed in the classification rates on these datasets, which were closer to 70% when compared with the ground truth provided by the subjects in these videos. In the case of the MIME dataset the classification rate obtained by the human annotators was in fact below the theoretical performance of a majority class recommender. In other words, a recommender always predicting the most common class in this dataset (positive valence) would obtain a higher accuracy than our crowd of annotators.

There are some remarks to be made here. Unlike many experiments, in our case the human annotators were volunteers who were told they could quit the experiment when they wished to do so. The ratings provided by the annotators who did not complete the entire experiment were then not taken into account, we reasoned that this way we would end up with only the more motivated annotators who completed the entire session, avoiding the users who gave up on their task after rating a few videos. We found that all of the remaining annotators completed the first three annotations of each dataset with perfect agreement, showing they had the skill necessary to complete the experiment. These two findings should ensure the quality of our annotator crowd.

Finally, in the case of the MIME experiment, the low classification accuracy might also be explained by the resolution of the recordings. The videos in this dataset were 144 x 144 pixels in size, of which the face of the subject was shown with around 60 x 75 pixels. There have been reports in literature stating that, for a human observer to properly recognize a face and the shown expression, there is a minimum resolution of around 100-200 pixels in width and height Pantic and Rothkrantz [2003], Campbell and Green [1965]. More recently, Du and Martinez [2011] conducted an experiment where human observers were shown static pictures of people posing with a certain emotion. They found that accurate labeling of the emotions by humans was possible with a resolution as low as 20 x 30 pixels. The dataset used in their experiment did however consist of photos of posing actors. On the other hand, the MIME dataset consists of moving pictures with subjects who showed naturally occurring subtle emotion.
6.3 Automatic and human labeling

This is likely to be more difficult to detect using low resolution footage.

**6.3 Automatic and human labeling**

The experiments we finally conducted to compare the results between automatic labeling methods and human labels provide interesting results (Figure 5.2). For each of the three evaluated datasets, we found that the automatic FEI method obtained a lower classification rate than our human annotators.

An additional explanation may be found in the use of accuracy to determine performance in these experiments. Our automatic methods were designed to try to detect both classes equally well, even for the unbalanced datasets. While this configuration resulted in high F1-scores, the classification rates do not reflect this, on the MIME dataset this even resulted in a classification rate below majority vote performance (Table 5.3). In practice, it would however be more valuable to have an algorithm that is able to detect infrequently seen classes, rather than being able to detect the most common class.

The performance gap between the human annotator performance and our automatic methods is different from the results published by researchers such as Janssen et al. [2013]. These researchers frequently found that their machine learning approaches obtained higher classification rates than human annotators. The gap between the results we found and the results reported in previous research may however in part be explained by the types of datasets that were used. In the experiment by Susskind et al. [2007] used a dataset of photos of people who posed with a certain emotion. In Janssen et al., the subjects were asked to recollected and describe emotional situations. This method of emotion elicitation is known to elicit strong emotions [Levenson et al., 1991], stronger than elicitation through the use of video footage.

Moreover, it is known that facial expression recognition software is better able to detect extreme expressions than subtle ones [Tkaličič et al., 2013]. Considering these automatic labeling methods rely on the use of automatically detected facial expressions, it is very well possible that humans are better able to pick up the subtle emotions shown by people watching video fragments, whereas automatic solutions are better able to distinct between the more extreme expressions.

A second difference with especially the research by Janssen et al. is the use of binary classification in our work. In their experiment, the annotators were asked to label video with one of five emotions, whereas in effect in our method only two labels were possible. The confusion matrices reported in their work show that the human annotators were frequently confused about happy and relaxed sessions as well as sad, angry and neutral sessions. In our work these labels would be grouped together, resulting in a higher number of correct classifications. No confusion matrices are reported for their automatic labeling experiments, making a comparison in a binary classification scenario not possible at this time.

The final difference between the two automatic classification methods is the choice of features. While our method is limited to the use of facial expressions, Janssen et al. used a more generic application to track various points in the faces of their subjects. In particular, this allowed them to include features such as the vertical orientation of the pupil of their subject, the ratio the eyes were visible (indirectly measuring blinking)
and the amount the mouth of the subject was opened. Neither of these features could be included using our data source. No performance figures are reported when their method was limited to the use of facial expression-like features, only the overall performance of their entire set of video-based features, which does not allow us to make a direct comparison between the two automatic methods.
Chapter 7

Conclusions & Future Work

In this chapter we provide the conclusion of our work as well as our conclusions in relation to the research questions we outlined in Chapter 1. Furthermore, we describe several directions that could be considered for future work on the classification of the emotion of TV viewers.

7.1 Conclusion

Now that we have completed our research, we can conclude with our main research question:

**Main research question** How can the classification of the emotion of people watching a video fragment be applied in a real-world setting when only using video recordings depicting the viewer?

We have seen in Section 5.3 that using our FEI method, it is possible to determine the valence of people watching video fragments with high performance. The results we found were better than the similar method seen in literature. We also discovered in our experiments with alternative training methods (Section 5.1.4), that these methods obtain similar performance figures when simulating cold-start situations. We also found that, with the exception of the most natural dataset (MIME), the obtained F1-scores were well above the performance of the baseline methods. Using the information obtained using our human annotation experiments, we did however find that the classification rate of our automatic methods remained below the human performance.

Overall we can conclude that using facial activity, detected using off-the-shelf commercial software, can be used for emotion classification of people watching video. Furthermore, it seems to be possible to develop a system that can do so in a cold start situation where no prior information is known about the user of the system. The performance is shown to be highest for datasets with high quality video recordings of the subject and rather extreme emotion elicitation videos, although even in less extreme situations the classification performance exceeds baseline performance.

In the next section, we will continue with the answers we found in relation to each of the research questions that supported our work.
7.2 Research questions

In order to find the answer to this main research question, seen in the previous section, we evaluated four sub research questions. These questions provided us with partial answers to our core question and we will further discuss these questions in this section.

**Research question 1** When the system has not seen a specific user before and is not familiar with their behavior, to what extent can methods found in literature be successfully applied?

In Chapter 5 we evaluated the performance of our implementation of the Onset Offset Counting method by Koelstra and Patras [2012], which is the main method we considered in our answer for this sub-question. We considered two different cross-validation methods for our machine learning algorithms. The first method was leave-one-session-out or personalized training, where the algorithm was trained to recognize the behavior of one specific user from the MAHNOB-HCI dataset [Soleymani et al., 2012] by training on 19 sessions of one user, after which the 20th session was used to test the classifier performance. The second method trained the machine learning algorithm in a more general way using leave-one-subject-out cross-validation. In this case, the algorithm was trained using the sessions for all but one subject and evaluated using the sessions of the last user.

When the first training method is used, there is prior knowledge about the user, while this is not the case with the general training method. We found that using generalized training resulted in a small penalty to the average F1-score of the classifications when compared to the personalized training. The accuracy was less affected by this and in some cases improved in the general case. In general this suggests that both the method found in literature as well as our own FEI method can be used in situations without prior knowledge about the user with similar performance levels.

**Research question 2** What is the influence of alternative data pre-processing and feature vector construction methods on classification performance?

To evaluate the impact of different data pre-processing and vector construction methods, we presented an alternative method (FEI) (Section 5.1.3). In this method we used different pre-processing and feature vector construction steps in comparison to a baseline method (OOC-SVM). We found that these alternative steps resulted in an improvement on the classification performance. While we only evaluated one such method, our results suggest that the classification performance can be improved by using different processing and aggregation steps.

Additionally we also evaluated the impact of two different feature vector optimization techniques, in particular recursive feature elimination (RFE) and feature scaling. We found that in general the use of feature scaling had a positive effect on the classification performance. The effect of using RFE was however more dependent on the training method and dataset used. When training specifically for one person, we the use of RFE improved the classification performance. Conversely, when training using information of a group of people, the use of RFE reduced classification performance.
**Research question 3** How well can facial expression-based emotion classification methods be used in uncontrolled environments, such as living rooms around the world?

To answer this third question we applied the OOC-SVM and FEI methods to three different datasets (Section 5.1.4). The first dataset considered (MAHNOB-HCI) was collected in a highly controlled lab environment, which did not match a living room situation. The second (BINED) and third (MIME) datasets did however contain recordings obtained in more natural settings. Our experiments showed that the extremeness of the videos had a larger effect on the performance of our method than the freer environment in which the subjects were observed.

The most natural dataset we considered, MIME, did however show that there are challenges to real-life applications of these methods. Due to the experimental setup in this dataset, the video quality obtained was in some cases insufficient for the facial expression recognition software to properly function. Furthermore the environment caused subjects to frequently move, which caused similar difficulties.

These challenges resulted in a number of sessions being discarded from this dataset. Despite these challenges, emotion classification turned out to be possible in these more natural environments. The F1-scores obtained by our method exceeded the performance of the baseline methods considered.

**Research question 4** How can we utilize people to determine a benchmark level of valence classification accuracy for emotion classification methods?

Due to the lack of baseline methods, we resorted to the use of human annotators to provide a baseline of sorts when it comes to classification performance. We developed a web-based video annotation tool, which allowed us to easily reach a large number of people who could aid our research with limited effort. In this annotation tool we first provided them with the instructions to complete the experiment followed by a small form to provide their demographic information. During the actual annotation we presented the user with a video and asked them to use a 5-point Likert scale to estimate the valence of the person shown in the video.

Using an increased video playback speed we managed to have a sizable number of videos annotated for each of the three datasets we evaluated in our work. To ensure a certain level of quality, we excluded all ratings from annotators who did not complete the entire experiment or who did not provide the expected answers to the first three annotation tasks. The resulting ratings showed an inter-rater agreement ranging from “substantial” to “almost perfect”, based on the Fleiss’ kappa values of up to 0.88. This shows that the use of human annotators is a feasible way to collect a target performance level that can be used to compare the performance of automatic classification methods on datasets that have not been used for classification purposes.

### 7.3 Future Work

During our research we have encountered several opportunities that can be considered in future research. We will discuss these opportunities in this section.
7.3 Future Work

7.3.1 Dataset collection

While there are many datasets that can be used for affective computing research, only a small number can be used for research on the topic of affective video recommender systems. Furthermore, the available datasets do not completely reproduce real-life situations.

In the ideal experiment, subjects would be recorded in their own homes while watching the stimulus video. This way, the subjects might be inclined to follow their regular routine, which could include behavior such as glancing at a smart phone or looking away from the television screen when the video shown is not interesting. When collecting a new dataset with video recommendation applications in mind, we suggest it would be good to ensure that there are ways for subjects to display their disinterest, which could be as simple as leaving a newspaper within reach of the subject.

Furthermore, careful attention should be paid to the stimulus videos that are chosen. In some datasets, the stimuli chosen were more extreme than one would normally encounter in most situations. These extreme stimuli result in very visible facial expressions. While these do allow for high classification rates, the similarity with real-life applications is limited. Example stimuli to select could include TV shows, news reports or YouTube videos. A second factor with the stimuli videos is the length of each session. In order to collect a large number of sessions, it is an option to limit the stimuli to one or two minutes of video, allowing for up to 60 sessions to be recorded per hour. By using shorter stimulus videos, the reactions that can be experienced by the subjects are also limited, as it is difficult to become bored with a video that lasts only a short moment.

When the goal is to develop affective recommender systems for elderly people, as was in part intended with the MIME dataset, special attention should be paid to the quality of the video recording. We experienced difficulty with the expression recognition software when used to recognize the expressions of elderly subjects, in particular when the subjects wore glasses. The use of high resolution recording equipment could reduce this problem, which could be resolved in future work by using a video camera with optical zoom.

7.3.2 Training techniques

An interesting direction for further research also exists in the area of training methods. In our work we evaluated the suitability of training for one specific subject with the information recorded from a group of other subjects. Future research could take this one step further by evaluating whether training on one dataset allows for classification of a second dataset. By doing so, it would be possible to evaluate whether changes in experimental environment, stimulus length and demographics significantly impact the classification performance of these methods, which is an important step towards the deployment of these technologies into our everyday lives.

A second possible variation would be to evaluate the suitability of these methods to use a “sliding window” classification of sorts. Currently these methods predict the rating of an entire session or the reported valence during a single video frame [McDuff et al., 2010], while in a real-world application it would be interesting to see if using for instance recordings of the previous five minutes can be used to determine whether
the user is still interested. This approach would allow for “interventions” in the form of recommendations to change to a different show once the user’s interest subsides, perhaps even before they are fully aware of this.

7.3.3 Data processing

In our work we evaluated the inclusion of features based on the subject’s heart rate, which has also been reported in various related works. In our evaluation the benefits seemed to be limited in comparison to the trouble of connecting ECG sensors. There are solutions on the market, such as Microsoft’s Kinect for Xbox One, that are able to detect the heart rate of people using only video recordings. Future research could evaluate the accuracy of these methods in comparison to “invasive” ECG measurements as well as the possible benefits to emotional classifications.

While there are a range of facial expression recognition applications and the impact of the recognition rate of the expression recognition software might play a big role, little is known about the relative performance of these methods. To the best of our knowledge, no independent comparisons between the academic state of the art and commercial solutions for FACS recognition have been published. The availability of this information would allow for a better comparison between published works that rely on different emotion recognition solutions.

7.3.4 Method evaluation

Finally there are still different scenarios to be evaluated using our method. It would for instance be interesting to evaluate the performance of our method on a dataset consisting of recordings for longer videos, such as news reports of several minutes or longer videos obtained from YouTube. Such evaluations might show whether the aggregation methods used still manage to result in sufficiently descriptive features to allow for accurate classification.

Another interesting option would be to conduct an evaluation where subjects are asked to watch longer video fragments while continuously rating their valence, as was done by McDuff et al. [2010] and evaluating the performance of our method in a sliding-window setting. This way, it could be evaluated whether our methods could be applied using a sliding window, where the system takes into account the behavior observed over the past minutes to determine whether or not the user is still enjoying the current video. Such an application would help with real-life applications, where “live” recommendations could change the way we watch TV.
Bibliography


Marko Tkalčič, Ante Odić, and Andrej Košir. The impact of weak ground truth and facial expressiveness on affect detection accuracy from time-continuous videos of facial expressions. *Information Sciences*, 2013.


Appendix A

Glossary

In this appendix we give an overview of frequently used terms and abbreviations.

**AU**: Action Unit, the most basic building block of a facial expression, according to FACS.

**BINED**: Belfast Induced Natural Emotion Database is a dataset of emotional displays.

**ECG**: Electrocardiography, recordings of the electrical activity of the heart.

**EEG**: Electroencephalography, recording of the electrical activity of the brain.

**EMFACS**: Emotional FACS, a subset of facial expressions from FACS, which are known to relate to certain emotions.

**FACS**: Facial Action Coding System, a system to classify facial expressions by describing basic movements in ones face, called Action Units.

**FACSAID**: FACS Affect Interpretation Dictionary, a dictionary of emotion-related facial actions.

**FEI**: Facial Expression Intensity, a method used for feature vector creation.

**HMM**: Hidden Markov model, a machine learning algorithm in which the classification task is described as a Markov process.

**HR**: Heart rate.

**HRV**: Heart rate variability, the variation in time between consecutive heart beats.

**MAHNOB-HCI**: Machine Analysis of Human Naturalistic Behavior: Human Computer Interfaces, a dataset of emotional displays.

**MIME**: Memorybanks Interactive Metadata Extraction, a TNO research project.

**OOC**: Onset/Offset Counting, a method used for feature vector creation.

**RBF kernel**: Radial basis function kernel is a kernel function used in SVM classification, which is able to transform the feature vectors to a higher dimensional space, sometimes improving the obtained classification performance.
**RMSSD:** Root mean square of successive differences, a method to calculate the HRV by calculating the root of the mean of the sum of the squares of differences between successive beat intervals.

**SVM:** Support Vector Machine, a machine learning algorithm.

**VAD:** Valence, arousal and dominance, a three-dimensional scale to describe emotional states.

**Valence:** The term used in the context of emotions to describe whether it is regarded to be positive (i.e. happiness) or negative (i.e. anger), also referred to as “pleasure”.
Appendix B

Dataset Exclusions

B.1 MAHNOB-HCI

Subjects 3, 9, 12, 15, 16 and 26 have at least one session without the full set of data and have been excluded.

B.2 BINED

The sessions from female subjects 20, 24, 28, 30, 38 and 42 and male subjects 8, 12, 13, 17, 25, 26, 28, 33, 35 and 42 were excluded due to no faces being detected or major artifacts in the received video files.

B.3 MIME

The sessions obtained for the last video stimulus watched could not be used due to missing self-assessment data or invalid video recordings for the following session ids:

- 6udngdmrfz85mi
- qkfm4gte64cuwhfr
- k3euidkki5m6lrx
- d0hbtv3y096br9
- 4q78xj7dazsyk3xr
- mhwdekyyd4zk6gvi
- yqc46y8ulahsemi
- pj4zgh8zo4bfn7b9
- 8y6fhtwzrozuxr

Finally the sessions rh1dolvgi90ms4i and zjqswbmlphf47vi were excluded because the Noldus FaceReader software could not detect a face in these video recordings.
Appendix C

Detailed Results

Certain results have been excluded from the body of this thesis for the sake of brevity. As these results might still be of value, we have included these tables in this appendix.

C.1 OOC-GNB evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Valence</th>
<th>Arousal</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>CR</td>
<td>F1</td>
</tr>
<tr>
<td>Koelstra and Patras [2012]</td>
<td>63.3</td>
<td>64.0</td>
<td>66.3</td>
</tr>
<tr>
<td>Our implementation</td>
<td>58.5</td>
<td>59.4</td>
<td>58.1</td>
</tr>
<tr>
<td>Random</td>
<td>48.7</td>
<td>50.0</td>
<td>48.1</td>
</tr>
<tr>
<td>Majority class</td>
<td>38.3</td>
<td>62.6</td>
<td>37.6</td>
</tr>
<tr>
<td>Class ratio</td>
<td>50.0</td>
<td>54.6</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Table C.1: Evaluation of the OOC-GNB method presented by Koelstra and Patras [2012] and our implementation of this method. Also shown are the performance figures of a random classifier and a classifier voting according to the majority class and the class ratios.
## C.2 Heart-rate evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Valence F1</th>
<th>Valence CR</th>
<th>Arousal F1</th>
<th>Arousal CR</th>
<th>Control F1</th>
<th>Control CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR-features</td>
<td>49.0</td>
<td>50.2</td>
<td>57.4</td>
<td>59.4</td>
<td>51.9</td>
<td>51.9</td>
</tr>
<tr>
<td>FEI</td>
<td>67.6</td>
<td>69.8</td>
<td><strong>60.7</strong></td>
<td><strong>62.7</strong></td>
<td><strong>61.3</strong></td>
<td><strong>61.5</strong></td>
</tr>
<tr>
<td>HR+FEI</td>
<td><strong>68.1</strong></td>
<td>63.5</td>
<td>58.1</td>
<td>59.8</td>
<td>60.4</td>
<td>60.6</td>
</tr>
</tbody>
</table>

(a) Personalized training on the heart rate features.

<table>
<thead>
<tr>
<th>Method</th>
<th>Valence F1</th>
<th>Valence CR</th>
<th>Arousal F1</th>
<th>Arousal CR</th>
<th>Control F1</th>
<th>Control CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HR-features</td>
<td>45.2</td>
<td>60.2</td>
<td>50.4</td>
<td>51.7</td>
<td>39.5</td>
<td>49.6</td>
</tr>
<tr>
<td>FEI</td>
<td><strong>64.8</strong></td>
<td>66.7</td>
<td><strong>74.2</strong></td>
<td>79.4</td>
<td>51.0</td>
<td>54.6</td>
</tr>
<tr>
<td>HR+FEI</td>
<td>62.0</td>
<td>63.5</td>
<td>48.4</td>
<td>50.2</td>
<td><strong>56.6</strong></td>
<td>56.7</td>
</tr>
</tbody>
</table>

(b) Global training on the heart rate features.

Table C.2: Classification using the heart rate features and fusion with FEI features, classification rates and F1-scores. The highest F1-scores per target are shown in bold.