Communicating Emotions:  
Multi-modal Emotion Recognition and The Experience of Emotional Intensity 

MSc Thesis 

Paul Tacken  
0413593 

Department of Artificial Intelligence  
Radboud University Nijmegen  

July 13, 2011 

Supervisors:  
Gert-Jan de Vries (Philips Research Eindhoven)  
Joris Janssen (Philips Research Eindhoven)  
Pim Haselager (Radboud University Nijmegen)  

Abstract

In the past years various studies have focussed on automated emotion recognition using various modalities (video, audio, physiology) separately. The work we present comprises an investigation of the usefulness of each of these separate modalities and various modality combinations, for automated emotion recognition. Video, audio, and physiology signals were recorded from persons telling about an emotional event from their life, for five separate emotions: happy, relaxed, sad, angry and neutral. Features were extracted for each modality separately. The data from the three modalities was merged using feature-level fusion. Classification was done for each modality separately and for every modality combination, using a Support-Vector Machine (SVM) and a MultiLayer Perceptron (MLP). When looking at the classification performances for each modality separately, physiology showed promising results, especially for the SVM classifier. For both classifiers, there was an overall improvement in classification performance when modalities were merged instead of used separately.

For automated emotion recognition, physiology appeared to be a quite influential modality. In order to investigate whether this could also be the case in human-human interaction, we conducted an experiment on the effect of a physiological signal on the judgment of the intensity of someone's emotional state. Participants were shown videos of persons telling about an emotional event (angry or neutral) from their life. Each video was accompanied by a heartbeat sound (slow or fast). Participants were told that the heartbeats were those of the persons in the videos at the moment they were telling the story. The participants were given the task to judge the emotional intensity of the persons the videos. Results showed that both faster heart rates and angry facial expressions caused higher emotional intensity ratings when compared to slower heart rates and neutral facial expressions. These results show the potential for a whole new way of communicating emotions.
Contents

1 General Introduction 3
   1.1 Contents of this thesis ........................................... 4

2 Multi-Modal Emotion Recognition 6
   2.1 Introduction .................................................... 6
   2.2 Methods .......................................................... 7
       2.2.1 Data gathering .............................................. 7
       2.2.2 Feature extraction .......................................... 9
       2.2.3 Fusion ....................................................... 14
       2.2.4 Feature selection .......................................... 14
       2.2.5 Classification .............................................. 15
   2.3 Results ............................................................ 16
       2.3.1 SVM Results ............................................... 16
       2.3.2 MLP Results ............................................... 20
   2.4 Discussion ...................................................... 22

3 Emotion Intensity Judgment 25
   3.1 Introduction .................................................... 25
       3.1.1 Research questions ......................................... 26
       3.1.2 Hypotheses .................................................. 26
   3.2 Methods .......................................................... 29
       3.2.1 Participants ................................................ 29
       3.2.2 Experimental design ....................................... 29
       3.2.3 Materials ................................................... 29
       3.2.4 Procedure ................................................... 31
   3.3 Results ............................................................ 32
   3.4 Discussion ...................................................... 34
       3.4.1 Influence of physiological signal .............................. 35
       3.4.2 Physiology in human-human interaction ....................... 36

4 General Discussion 37
   4.1 Summary .......................................................... 37
   4.2 Applications ..................................................... 38
   4.3 Future work ...................................................... 38
   4.4 Conclusion ...................................................... 39

References 40
A Data gathering 43
   A.1 Participants and Design 43
   A.2 Materials 43
   A.3 Procedure 44

B List of extracted facial features 46

C Instruction form of the Emotional Intensity Judgment experiment 47

D Adapted Multi-Dimensional Emotional Empathy Scale 49
Chapter 1

General Introduction

More than we probably realize it, emotions constantly influence our behavior during our daily lives. As Picard (1995) notes, our emotions, and those of others, determine for a great deal how we act upon given situations. Emotions can be describes as subjective experiences that are responses to particular events. They reveal what is important to us and what we are concerned about (Frijda, 1988). The events from which emotions are triggered can be internal (i.e. a thought) or external (Salovey & Mayer, 1990), and their consequences for the person determine the nature of the emotion that is being triggered (e.g. whether they are positive or negative). Emotions are subjective, because they are experienced differently from person to person. Salovey and Mayer (1990) state that proper management of one’s own emotions and the recognition of others’ emotions can be used to guide thoughts and actions. The ability to do this falls under the category of what they call emotional intelligence.

Although it may not seem very intuitive, there is reason to believe that emotions can support rational thinking. In fact, Damasio (1994a, 1994b) states that for rational thinking, emotion is indispensable. When faced with decisions, emotions can be used to confine the space of possible decisions to those that will give the most advantageous outcomes. Emotions can be used to ‘mark’ certain types of outcomes to certain types of behaviors as positive or as negative in order to create a bias for the behaviors that result in more advantageous outcomes. The purpose of these marks is to confine the decision space so that for future events profitable decisions can be made in a small amount of time. Humans don’t just have emotions, they also express (i.e. communicate) them. As noted by Frith (2009), the communication of emotions often happens unconsciously when interacting with others. Humans express emotions through various modalities, such as facial and vocal expressions, body posture, and physiology. The expression of emotions serves several purposes, to name a few examples: they can communicate some internal state (e.g. frustration), so that others can take this into account when interacting with the person who experiences this emotion (e.g. “don’t bother him/her, unless it is really important”); they enable one to let others know how one feels about certain issues that come to one’s attention (e.g. show approval/disapproval); and they can be used to transfer an emotional state to another persons (e.g. cheer up someone who is sad). However, emotional expressions can also be used to deceive people, in case the displayed emotional expressions does not represent one’s actual emotional state (e.g. when telling a lie), as argued by Ekman (2009).

According to several researchers, (e.g., Picard, Vyzas, & Healey, 2001; Sebe, Cohen, & Huang, 2005), people are capable of reading other peoples’ affective states through, not only facial and vocal expressions, but also their physiology. Examples of such signals are the (hand) sweatiness, heart rate and respiration. The ability to recognize and reckon with other peoples’ affective states
is essential for successful human-human interaction. Moreover, emotional skills are considered to be an important component of overall intelligence (Pantic & Rothkrantz, 2003; Picard et al., 2001; Sebe et al., 2005).

The link between emotion and intelligence does not only apply to humans (and animals), it also applies to computers:

Machines may never need all of the emotional skills that people need but they will inevitably require some of these skills to appear intelligent when interacting with people. It is argued that to truly achieve effective human-computer intelligent interaction (HCII), there is a need for the computer to be able to interact naturally with the user, similar to the way human-human interaction takes place. (Sebe et al., 2005)

In the past two decades, the field of Affective Computing has grown to become a field of much interest. Picard (1995), who is regarded as the pioneer in this field, defined affective computing as “computing that relates to, arises from, or influences emotions”. Many, including Picard, have argued that no intelligent human-computer interaction is possible if the computer is not able to take into account the human’s affective state (e.g., Fairclough, 2009; Pantic & Rothkrantz, 2003; Picard et al., 2001; Sebe, Lew, & Huang, 2004).

A computer that could recognize and appropriately respond to a user’s affective states could significantly improve the HCI experience for users. One could think of applications that are able to detect stress and which can for example, suggest to take a break, or play relaxing music. Furthermore, a computer’s ability to recognize a user’s affective state could also improve remote communication between humans by transferring information about the other person’s affective state, which would be naturally perceived in a face-to-face interaction.

As already mentioned above, humans express emotions through multiple modalities (e.g. facial and vocal expressions, body posture, and physiology). Most studies on automated emotions recognition so far have focused on facial and vocal signals. A growing number of studies is now also starting to focus on physiological information. Physiological signals have been shown to be very useful for separating emotions (e.g., Picard et al., 2001). Several researchers (J. Kim & André, 2008; K. Kim, Bang, & Kim, 2004) argue that reading emotions from physiology is more natural than emotion recognition from facial or vocal information, since physiological signals, which represent the activity in the autonomic nervous system (ANS), more directly reflect an emotional state, whereas facial and vocal expressions as the result of an emotion can more easily be repressed. In this thesis the usefulness of each of these modalities for automated emotion recognition and the judgment of emotional intensity is investigated.

1.1 Contents of this thesis

Sebe et al. (2004) point out that (1) if in a human-human interaction one human would be replaced by a computer, the affective interaction disappears, since a computer than generally does not take the user’s affective state into account. Also, (2) if a face-to-face communication of two persons would be replaced by communication via computers (e.g. through e-mail or instant messaging), then much affective information would disappear. These points narrow down the motivation for the research presented in this thesis quite well. In this thesis, two research topics are described. The first topic is in the area of multi-modal emotion recognition, and consists of the following research question:

1. Which modalities and which modality combinations would be most suitable to be used for
automated emotion recognition?

In order to answer this research question, we investigated multi-modal emotion recognition, using data from five emotional states. First, data was gathered from five emotion classes: happy, relaxed, sad, angry and neutral. Autobiographical recollection is used as the method for eliciting emotions; participants are telling about emotional events from their life, that have been selected so that they each represent one the five emotions indicated above. Data is gathered from three modalities: video (facial expressions), audio (vocal expressions) and physiology. In order to be able to apply multi-modal emotion recognition, the information from the three modalities needs to be combined. For this, we use a fusion method called feature-level fusion, by which the features from different modalities are concatenated. Before the features are used for classification, feature selection is applied in order to find the most useful features and to filter out the ones that cause noise. Classification is done using two classifiers, a Support-Vector Machine (SVM) and a Multi-Layer Perceptron (MLP). Three classification methods are applied, Leave-one-Out (LVO): training on the data of all but one subjects and testing on the left-out subject; Cross Validation per Subject (CVS): apply cross validation to the data of all subjects separately; and Cross Validation Overall (CVO): apply cross validation to the combined data of all subjects. The results will reveal which modality and which modality combination are most successful for automated emotion recognition, at least for the kind of data we use.

The second research topic is about the judgment of emotional intensity by humans and tries to answer the following research question:

2. Can the judgment of someone’s emotional intensity be significantly influenced by a physiological signal?

In order to answer this research question, an experiment is performed in which participants are given the task to judge another person’s experienced emotional intensity, based on their facial expressions and a physiological signal, i.e. a heartbeat. An experiment is performed, in which participants watch videos of persons telling about an emotional event from their life. The event is either a neutral one (e.g. going home) or one where the person experienced anger. The audio was muted so that participants are not able to hear the story of the person in the video. During each video, an artificial (slow or fast) heartbeat sound is played, of which participants are told that it is the heartbeat of the person in the video. The participants’ task is to judge the emotional intensity of the person in the video. Our goal is to investigate the influence of the physiological signal in case it is in agreement with, or contradicts the facial expressions of the person in the video. Also, using a questionnaire, the participant’s experience of hearing the heartbeat sound is investigated, by asking them how they felt about hearing the heartbeat sound.

Our research on multi-modal emotion recognition is described in Chapter 2. This is followed by our research on the judgment of emotional intensity, which is described in Chapter 3. Finally, in Chapter 4 a discussion is presented regarding the results of the presented work, as well as recommendations for future work.
Chapter 2

Multi-Modal Emotion Recognition

2.1 Introduction

Automated emotion recognition is a relatively young field of investigation. Studies have been done on automated emotion recognition based on facial expressions (e.g., Ahn, Bailenson, Fox, & Jabon, 2009; Michel & El Kaliouby, 2003; Sun, Sebe, Lew, & Gevers, 2004), vocal expressions (e.g., Schuller, Lang, & Rigoll, 2002) and physiological signals (e.g., Frantzidis et al., 2010; J. Kim & André, 2008; K. Kim et al., 2004). For gathering emotion data, many studies use deliberate displays of facial and/or vocal expressions, instead of spontaneous (“real”) expressions (e.g., Michel & El Kaliouby, 2003; Nogueiras, Moreno, Bonafonte, & Mariano, 2001; Sebe, Cohen, Gevers, & Huang, 2006). As pointed out by Pantic (2009), systems that are trained on deliberately displayed expressions, will probably not do well in real-world situations, since the deliberate expressions are often exaggerations of expressions as they occur in natural settings. Sun et al. (2004) claim to be first to have created an authentic emotion expression database. The emotional expressions were elicited using a kiosk equipped with a hidden camera, playing parts of movie trailers from different genres. This approach has the advantage of being located outside of a laboratory setting, so that people would feel less inhibited to show their true emotions.

Several approaches have been used for extracting facial features. Examples are the use of local binary patterns (LBP) for creating texture descriptions of the face area (e.g., Shan, Gong, & McOwan, 2009), automated tracking the position of specific facial feature points (eyebrows, eye corners and mouth corners) in the face area (e.g., Ahn et al., 2009), and tracking local deformations of facial features based on a 3D wireframe model of the face (e.g., Sebe et al., 2006). According to Zeng, Pantic, Roisman, and Huang (2009) most of the currently existing facial expression recognition methods are based on 2D spatio-temporal features.

For expression recognition from speech, mainly pitch and energy appear to be among the most popular vocal features (e.g., T. S. Huang, Chen, Tao, Miyasato, & Nakatsu, 1998; Nogueiras et al., 2001; Van den Broek, Van der Sluis, & Dijkstra, 2009). Few attempts have been made to integrate also semantic information from a speech signal (e.g., Schuller et al., 2002). For physiology, often used features are electrodermal activity (EDA), electrocardiogram (ECG) and skin temperature (ST) (e.g., K. Kim et al., 2004; Nasoz, Lisetti, Alvarez, & Finkelstein, 2003), as well as respiration (RSP) (e.g., J. Kim & André, 2008).
More and more studies are starting to focus on emotion recognition based on combined information from multiple modalities. Combining information from multiple modalities has been shown to improve classification performance over using information from a single modality (Busso et al., 2004; T. S. Huang et al., 1998). Most of the current studies on multi-modal emotion recognition use a combination of facial and vocal expressions (e.g., Sebe et al., 2006; Zeng, Zhang, Pianfetti, Tu, & Huang, 2005). However, there have also been studies combining other modalities, such as video and physiology (e.g., Bailenson et al., 2008). Several methods exist for combining (“fusing”) the information coming from multiple modalities. These can be divided into feature-level fusion (“early fusion”), decision-level fusion (“late fusion”) and model-level fusion. In decision-level fusion, the data is classified for each modality separately, after which the classification results are combined. As noted by e.g., Sebe et al. (2005) and Zeng et al. (2009), in this fusion method, the implicit assumption is made that the signals from different modalities are independent, which is clearly not the case. In feature-level fusion, the feature vectors of the modalities that are to be combined are concatenated, creating combined feature vectors. For this method it is necessary that timing information of the modalities matches exactly. This is not necessarily the case in model-level fusion, where the information from different modalities is more loosely connected, through the use of, for example, probabilistic graphical models such as hidden Markov models, Bayesian networks and dynamic Bayesian networks, as described by Sebe et al. (2006) and Zeng et al. (2009).

In the remainder of this chapter we will explain our research on multi-modal emotion recognition. The purpose of this research is to see which modality and which modality combination would be most useful for automated emotion recognition. For this, data from five emotion classes is gathered, these are: happy, relaxed, sad, angry and neutral. The modalities used are video (facial expressions), audio (vocal expressions) and physiology. Three classification methods are applied, Leave-one-Out (LVO), Cross Validation per Subject (CVS) and Cross Validation Overall (CVO). These methods are evaluated using two classifiers, a Support-Vector Machine (SVM) and a Multi-Layer Perceptron (MLP). Furthermore, we investigated the claim by Wilder (1962) that arousal (how active an emotion is) is best predicted by physiology whereas valence (how positive or negative an emotion is) is best predicted by facial expressions. The methods that were used for data gathering, feature extraction, fusion, feature selection and finally classification are described in the next section.

2.2 Methods
The methods utilized for automated emotion recognition are explained in this section. These can be categorized into several steps. First, video-, audio-, and physiological data was gathered from persons experiencing a pre-defined set of emotions. Second, features were extracted for each modality separately, using different software tools for each modality. Third, modalities were fused using feature-level fusion. Fourth, feature selection was applied in order to find the most useful features for separating the emotion classes. Finally, classification was done for each modality separately and for every modality combination, using a Support-Vector Machine (SVM) and a Multi-Layer Perceptron (MLP). In Figure 2.1, an illustration of the process is given.

2.2.1 Data gathering
The data gathering process was performed before start of our research, by Joris Janssen. Below a brief description of this process is given. For a full description of the data gathering process we direct the reader to Appendix A.
Humans are capable of having a great number of diverse emotions. For this research project on emotion classification a selection needed to be made. For this, the valence/arousal model by Russell (1980) was used. The five selected emotions were happy, relaxed, sad, angry and neutral. Each emotion, except neutral, represents one of the four quadrants of the model. Neutral is located in the middle. See Figure 2.2 for an illustration of this.

Emotion elicitation

For this research project the authenticity of the expressed emotions was essential in order for the physiological measurements to be of any use. Hence, acted expressions (facial or vocal) are of no use here. For obtaining reliable video-, audio- and physiological data of a persons emotion sensation, the emotions should be felt and actively expressed in laboratory setting as they would be in daily life. In order for the elicited emotions in this experiment to be as real as possible, autobiographical recollection was used. In this method participants are telling about an emotional event from their life when a particular emotion was experienced. This approach of eliciting emotions has been shown to be quite effective (e.g., Schaefer & Philippot, 2005).

Procedure

Video-, audio-, and physiological data was gathered from 17 subjects telling about an emotional event from their life, for five emotions: happy, relaxed, sad, angry and neutral. The data from one subject was removed, because of too short story durations. For each of the five selected emotions subjects were instructed to talk about an event in their life where this emotion was
experienced by him/her (the autobiographical recollection task). Each story was preceded by a five-minute baseline phase, in which a calming video is shown, and a preparation phase. In the preparation phase subjects could mentally relive the situation where they experienced the specified emotion. After the autobiographical recollection task, subjects were instructed to look back at their own recorded video, and to continuously rate their experienced arousal and valence on the scale of 1 to 9. The video data was gathered from the frontal view of the test subjects, looking almost directly into the camera. From the resulting videos the parts were cut in which participants were talking, leaving out the baseline- and preparation parts. These video parts had a resolution of 320x240 pixels. A microphone was used for recording speech. The physiological data was gathered using a Nexus-Recorder. The physiological signals that were recorded are Skin Conductance (SC), Skin temperature (Temp), Respiration (RSP) and Electrocardiogram (ECG), with recording frequencies of 128 Hz for SC, Temp and RSP and 1024 Hz for ECG.

2.2.2 Feature extraction

Time window

In order to extract feature values from the gathered data, an appropriate time window needed to be selected. We used a time window for describing the duration of the video/audio/physiology signal over which data values are averaged, so they could be used as feature values. Several options for the duration of the time window were considered. In order to get many features, a time window of 1 second — shifting 1 second each time — would be sufficient. However, physiological signals have a few seconds delay, compared to facial and vocal signals with regard to the response time to an emotion sensation. When combining information coming from different modalities, the feature values of both modalities should — approximately — correspond to the same moment in time when the emotion was felt. In order to do a reliable matching, 1 second is too short. Also, feature values averaged over a time window of only 1 second would result in large interpersonal differences, and they would be harder to generalize to certain emotions. In order to increase robustness of the matching process, and to improve generalizability of the feature values, while keeping the number of resulting feature vectors large enough to do robust training and testing, a time window of 10 seconds was selected.

Time step

Several time step options have been explored. In case of a time window of 10 seconds, the number of feature values could be increased by having the time windows overlap. For first tests, a time step of 1 second was chosen, causing the time windows to overlap 9 seconds. However, the problem with this overlap is that, when splitting the data up into a training set and a test set, the test data is not independent of the train data, causing the classification results to be artificially high. For this reason, we decided that the time windows should have no overlap. Therefore, a time step of equal length to the time window (10 seconds) was selected.

Facial features

For extracting facial features from the video files, two methods have been taken into consideration. One method was to use local binary patterns (LBP) — first introduced by Ojala, Pietikäinen, and Harwood (1996) — of facial images as features for facial expression recognition. Shan et al. (2009) describe the creation of LBPs in several steps. The first step is to place a grid over the facial image (after the facial image has been found), dividing it into several regions, so that
each region contains an equal amount of pixels. Then, for each pixel a binary code is generated, describing its relation to all its neighboring pixels. See Figure 2.3 for an example of this process.

![Figure 2.3: Creation of binary code for each pixel. Adopted from (Ahonen et al., 2004).](image)

Next, for each region, a histogram is created describing the pixel values in that region. This histogram is called the LBP histogram. Concatenating all LBP histograms will result in a feature histogram of the facial image (Figure 2.4). Such feature histograms can subsequently be used for emotion classification by means of, for example, template matching. For an elaboration of the use of LBPs for facial expression recognition, we refer the reader to (Shan et al., 2009).

![Figure 2.4: Concatenation of LBP histograms from every region of the face area. Adopted from (Shan et al., 2009).](image)

The other method that was taken into consideration consists of tracking the position of specific facial feature points, e.g., eyebrows, eye corners and mouth corners, in the face area, and to use the positions of these points as feature values, as was done by for example Ahn et al. (2009). See Figure 2.5 for an example. Ahn et al. (2009) used the OKAO Vision software package, which could automatically find and track, not only facial feature points, but also features such as the pitch, yaw and roll positions of the face, as well as the eye and mouth openness levels.

Although both methods appeared promising, for the sake of the feature vector length, being much greater for the former instead of the latter method, the latter method was selected. Using the OKAO Vision software package, 43 features were extracted, with an extraction rate of 15 feature points per second. The extracted feature points were then averaged for every 10 seconds of data. The complete list of extracted facial features is shown in Appendix B.

Vocal features

The selection of features that were used for this study is based on their potential link to emotion, found in previous studies on this topic. In literature on automated emotion recognition pitch is often mentioned as a useful feature. Pitch, many times referred to as fundamental frequency $f_0$, represents vocal-cord vibration as a function of time (Boersma & Weenink, 2010). Pitch can, for
example, be used to distinguish happiness or sadness from a neutral state (Cowie & Douglas-Cowie, 1996). Intensity is a popular feature as well. It represents the speech signal intensity in dB (Boersma & Weenink, 2010), and can be used for detecting for example anger, fear or happiness. Formants describe vocal resonances at specific frequencies in the frequency spectrum. Each formant has its own frequency range. The use of formants has also been found useful for the emotion classification (Ververidis & Kotropoulos, 2006). Nygaard, Patel, and Queen (2002) linked sadness to increased jitter and shimmer values. Jitter is a term for the frequency variability of the pitch signal and shimmer describes the amplitude variability of the pitch signal (Z. Huang, Chen, & Harper, 2006). The fraction of unvoiced parts in the pitch signal was also used as a feature value — as was also done by e.g., Busso et al. (2004) — in which each part has a duration of 0.01 seconds. Finally, the degree of voice breaks — which is the relative duration of breaks between the voiced parts (Boersma & Weenink, 2010) — was used.

For extracting audio features, the software tool PRAAT (Boersma & Weenink, 2010) was used. The extracted features were pitch: mean and standard deviation; intensity: mean and standard deviation; formants F1 to F4; jitter; shimmer; the fraction unvoiced frames and the degree of voice breaks. In total, 12 audio features were extracted.

**Physiology features**

A list of physiology features most commonly used in physiology based emotion recognition studies is given by Van den Broek, Janssen, Westerink, and Healey (2009). Among these are skin conductance (SC), which is a term for describing the electrical conductivity of the skin as a result of the clamminess of the skins surface; skin temperature (Temp), which is the temperature measure at the skin surface; respiration (RSP), which describes the breathing behavior; and electrocardiogram (ECG) features.

For extracting these features, a MATLAB implementation developed within Philips Research Eindhoven was used, which enabled the extraction of average feature values over a given time window. For all but the ECG signal, a time window of 10 seconds was used. In order to get
reliable ECG features, a time window of at least 30 seconds was necessary. Hence, a 30-second time window was used for the HR signal, which shifted with time steps of 10 seconds, causing subsequent time windows to have a 20 seconds overlap. The overlap is necessary in order to let the number of features match the other physiology features. The following features were extracted:

- SC: mean, standard deviation, slope and number per second.
- Temp: mean, standard deviation and slope.
- RSP: breathing rate in breaths per minute.
- ECG: mean, low frequency power, high frequency power, low/high frequency power ratio, RMSSD (root mean square successive difference) and SDNN (standard deviation of NN intervals).

A brief description of each term is given in Table 2.1.

Table 2.1: Brief description of the extracted physiology features

<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Average signal value over the given time window</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>Standard deviation over the given time window</td>
</tr>
<tr>
<td>Slope</td>
<td>(Positive or negative) incline of the linear fit computed over the raw signal</td>
</tr>
<tr>
<td>(SC) Number per second</td>
<td>Number of Skin Conductance Responses (SCRs) per second</td>
</tr>
<tr>
<td>Breathing rate</td>
<td>Number of breaths per 60 seconds</td>
</tr>
<tr>
<td>Low frequency power</td>
<td>Frequencies from 0.04 to 0.15 Hz, representing interbeat interval (IBI)** oscillations.</td>
</tr>
<tr>
<td>High frequency power</td>
<td>Frequencies from 0.15 to 0.4 Hz, representing the respiratory frequency band. (Berntson et al., 1997)</td>
</tr>
<tr>
<td>low/high frequency power ratio</td>
<td>Proportion low to high frequency, given by (power low)/(power high)</td>
</tr>
<tr>
<td>RMSSD (Root Mean Square of Successive Differences)</td>
<td>Root mean square of successive differences between IBIs**</td>
</tr>
<tr>
<td>SDNN (Standard Deviation of NN (normal to normal) intervals)</td>
<td>Standard deviation of the IBIs** within the time window</td>
</tr>
</tbody>
</table>

**IBI stands for InterBeat Interval, a term for describing the time period between successive heartbeats.

Class labels

For all modalities, the feature vectors of each class were labeled according to the emotion category of the story told by the test subjects. Before each story, there was a preparation phase in which subjects could mentally relive the situation where they experienced the specified emotion, so that the sensation of the specified emotion was present from the beginning of the story telling phase. The presence of the emotional sensation during the story telling parts was checked using the continuous valence/arousal ratings that were given by subjects for their experienced emotion during the story (see Section 2.2.1). The stories belonging to happy and relaxed were expected
to have higher average valence ratings than angry and sad. Likewise, arousal was expected to be higher for happy and angry than for relaxed and sad. For both valence and arousal, neutral was expected to be somewhere in the middle. Figure 2.6 shows the position of each of the five emotions in the valence/arousal space, according to the subjects’ ratings. If we would draw a horizontal and a vertical line through the point where neutral is located, we can see that the relative position of each emotion, except for sad, matches the quadrant it belongs to according to the valence/arousal scale, depicted in Figure 2.2.

Figure 2.6: The five emotional states positioned on the valence/arousal space according to the continuous valence- and arousal ratings given by the subjects.

Baseline correction

For physiological signals, a linear trend is known to occur in both mean and standard deviation. As a result, different signal values for different stories could be caused, not only by the emotions linked to the stories, but also by differences in time when the stories were told. Whether changes in physiological signals are actually caused by different emotion, or whether - according to James (1890) - different emotions are caused by changes in physiology, will be left an open question, as it falls outside of the scope and aims of this project.

In order to filter out the time factor, baseline correction was applied. At first, the entire baseline value was used for baseline correction. Later on, it was decided that this may not be the best approach. As explained in Section 2.2.1, the data gathering experiment consists of five parts, each of which consists of a baseline measurement, a preparation phase and a story telling phase. After the story telling phase of one part — in which some emotion was experienced — had ended, it was immediately followed by the baseline measurement of the next part. The signal values measured in the first part of the five-minute baseline period were likely to carry the effects of the previously felt emotion. Therefore, only the last 90 second-part of the baseline was used for correction. The applied correction method is described by following equation:
\[ x' = \frac{x - \mu_b}{\sigma_b} \]  

(2.1)

Where \( x \) is the original feature value; \( \mu_b \) is the baseline mean; \( \sigma_b \) is the baseline standard deviation; and \( x' \) is the newly computed feature value.

Scaling

Before the features can be used for classification, they first need to be scaled. There are two reasons for this. The first reason is that scaling prevents features in a large numeric range from dominating features in a small numeric range (Hsu, Chang, & Lin, 2003). This is necessary since the value range varies greatly among features. The other reason is that scaling removes the interpersonal range differences, since scaling is applied on the data from each person separately. This improves a classifiers ability to generalize over data coming from different persons. Two scaling methods were considered. The first method was to — for each person — scale all features between -1 and 1, as proposed by Hsu et al. (2003). The minimum feature value is scaled to -1 and the maximum value to 1. All other feature values are scaled between them, such that the original relation between the feature values is the same as before. The other scaling method was to compute z-scores and to use these as feature values. Z-scores are computed as follows:

\[ z = \frac{x - \mu}{\sigma} \]  

(2.2)

Where \( x \) is the original feature value; \( \mu \) is the average feature value; \( \sigma \) is the standard deviation over all feature values; and \( z \) is the z-score that is used as the new feature value.

Since this scaling method does not determine a fixed range for the feature values, this range can vary greatly from person to person. Not surprisingly, first results showed that classification rates for the features scaled with the z-scores were much worse than those of the features scaled between -1 and 1. Hence, the \{-1:1\} scaling method was applied.

2.2.3 Fusion

The recording-times of the physiology features were registered relative to the start of the video-recording, as were the recording times of the extracted video/audio features. Using this information, feature-level fusion (concatenating the feature-vectors from multiple modalities) was applied, whereby different feature vectors could be matched with great precision. Applying this fusion method enabled the use of several feature selection methods, which are described in the next subsection.

2.2.4 Feature selection

Not all features contribute equally to dividing the different emotion classes. Some features cause noise and bring down classification performance. In order to find the features that best separate the five emotion classes, feature selection was applied. Two feature selection methods were used. The first method was to use a wrapper approach for finding the best feature subset. A wrapper is constructed of a search algorithm for finding the optimal subset and an evaluation algorithm for evaluating the quality of each subset (John, Kohavi, & Pfleger, 1994). A Support-Vector Machine (SVM) and a MultiLayer Perceptron (MLP) classifier were used as evaluation algorithms. The search algorithms that were selected are best first search and genetic search. For both classifiers,
the two search algorithms were separately tested. For this method, the open source software tool WEKA (Hall et al., 2009) was used.

The other method consisted of making a ranked list, indicating how much each feature would contribute to separating the emotion classes. The choice was made to use the Chi-square ranking method for this, also using WEKA, which is analogous to the approach used by Bailenson et al. (2008) for ranking combined video- and physiology features. Then, for this ranked list, an iterative process was executed, in which the last ("worst") feature of the ranked list was removed, and the new ranked list was evaluated, until there were no more features in the list.

2.2.5 Classification

Classifiers

Two classifiers were used, an SVM and an MLP. An SVM is a classifier which uses hyperplanes to linearly separate data from different classes in a multidimensional feature space. Several kernel functions can be used, which are the linear, polynomial, radial basis function (RBF) and sigmoid kernel. Hsu et al. (2003) propose to use the RBF kernel, since it has several advantages over the other kernels: it is capable of separating data that is not linearly separable, its complexity is relatively low and it has less numerical complications. An MLP is a type of feedforward neural network which consists of an input layer, one or more hidden layers and an output layer. The nodes of each layer are connected to every node of each neighboring layer through weights. This is referred to as "full connectedness". The input value of a node (e.g. node n) is the sum of the output values of each node in the previous layer, multiplied by the weight of their connection to node n. Each node has an activation function, which transforms the input value to a new value, which is then given as the output of this node. When presented with train data, the MLP learns by adjusting the weights between the nodes until a desired input-output mapping is reached. Like the SVM, the MLP is good at dealing with data that is not linearly separable, and also does it have good generalization properties when fed with new data (Gardner & Dorling, 1998).

For both classifiers, no attempt has been made to optimize the parameter settings for better classification results, since the purpose of this research study was to investigate how successful certain modality combinations would be for separating a given set of emotion classes, not to find the best possible classification results. However, one parameter adjustment was made for the SVM: the cost and gamma values were changed to 100 and 0.1, respectively. Pilot tests showed that these gave considerably higher performance rates when compared to the default values.

Classification methods

Classification was done for each modality separately as well as for every modality combination. we applied three classification methods: Leave-one-Out (LVO), Cross Validation per Subject (CVS) and Cross Validation Overall (CVO). In LVO, the data of all but one subject is used for training, after which the data of the left out subject is used for testing. In CVS, training and testing is done for each subject separately, using 5-fold cross validation. Finally, in CVO, 5-fold cross validation is applied to the combined data of all subjects.

The CVO method is seen as the most meaningful here, since this method of testing the ability of classifying emotions seems most analogous to the way humans are “trained and tested” in classifying emotions. During their life, humans gain much experience in the recognition of emotions by being in contract with a numerous amount of people. When encountering a new person, the so far gathered experience (“training”) from all previous encounters can be applied in order to successfully determine the affective state of this new person (“testing”). However, the number of subjects in this research is quite small (N = 16) too small to make any solid
generalizations. Therefore, the classification method CVO, in which the train and test data originate from the same set of subjects, instead of LVO, is considered here as the best comparison to human emotion recognition of a new person.

The SVM classifier was applied to all three classification methods, but since the training and testing process is much more time consuming for the MLP classifier, and since the emphasis in this research is on the CVO results, the MLP classifier was only applied to CVO.

Valence/arousal classes

There appear to be indications that differences in valence are best detected by ones facial expressions, whereas arousal differences appear to be best detected by physiological signals (Wilder, 1962). The question is whether this research could be used to confirm this. During the data gathering experiment, subjects rated their felt valence and arousal after each telling phase, on a scale from 1 to 9. These valence and arousal ratings were used to re-label the emotion classes.

Three methods were used. In the first method, the valence and arousal values were used to re-compute the emotion classes. This was done as follows: if a valence score was $\geq 5$, it was considered positive, and otherwise negative. Analogously, if an arousal score was $\geq 5$, it was considered active, otherwise it was considered passive. Based on the valence/arousal model shown in the Section 2.2.1, emotion labels were assigned corresponding to the matching quadrant in this model (e.g. low valence + high arousal = angry). Only high and low valence/arousal values were considered, leading to 4 emotion classes. Hence, in this method the neutral class is not used. In the second method, the emotion classes were changed to valence classes, again by using the overall valence ratings, leading to two classes: positive and negative valence. Finally, in the third method, the emotion classes were changed to arousal classes, by using the overall arousal ratings, leading to two arousal classes: active and passive. The purpose of the second and third method was to test whether it was indeed the case that arousal could best be predicted by physiology and valence could best be predicted by facial expressions.

2.3 Results

In this section the classification results for the Support-Vector Machine (SVM) and the Multi-Layer Perceptron (MLP) are presented. For both classifiers, the main results are first presented, followed by the results for the relabeled data using the valence and arousal ratings. First, the results for the SVM classifier are explained, followed by the results for the MLP classifier.

2.3.1 SVM Results

Table 2.2 describes the classification results for each modality separately and for every modality combination. The results for the three classification methods Leave-One-Out (LVO), Cross Validation per Subject (CVS) and Cross Validation Overall (CVO) are separately shown. In Figure 2.7 the CVO results for the SVM classifier are visualized.

The LVO results represent the outcome of, first, for all (n = 16) subjects, training on the data of n-1 subjects and testing on the left-out subject, followed by taking the average over the performances. The highest LVO performance was gained with the modality combination of video and audio, with a classification rate of 30.36%. The CVS results are the averages of carrying out 5-fold cross validation over the data over the data of every of the 16 subjects separately, and then taking the average. Physiology by itself was the modality that gave the best classification performance, with a classification rate of 92.70%. Finally, the CVO classification results were
Table 2.2: An overview of the classification results for the SVM classifier, for every modality combination and three classification methods, Leave-One-Out (LVO), Cross Validation per Subject (CVS) and Cross Validation Overall (CVO).

<table>
<thead>
<tr>
<th></th>
<th>Physio.</th>
<th>Video</th>
<th>Audio</th>
<th>LVO</th>
<th>CVS</th>
<th>CVO</th>
</tr>
</thead>
<tbody>
<tr>
<td>NO FEATURE SELECTION</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x</td>
<td>25.40%</td>
<td>92.70%</td>
<td>75.94%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>26.30%</td>
<td>76.22%</td>
<td>58.59%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>27.78%</td>
<td>47.96%</td>
<td>39.07%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>28.78%</td>
<td>86.17%</td>
<td>79.30%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>25.36%</td>
<td>87.88%</td>
<td>72.45%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>30.36%</td>
<td>77.09%</td>
<td>64.57%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>28.78%</td>
<td>83.93%</td>
<td>79.30%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENETIC SEARCH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x x</td>
<td>90.39%</td>
<td>81.40%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEST FIRST SEARCH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>92.70%</td>
<td>82.24%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHI SQUARE RANKING</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>x x</td>
<td>33.63%</td>
<td>91.99%</td>
<td>81.12%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

obtained by doing 5-fold cross validation over the dataset of all subjects taken together. For CVO, the classification performances that resulted from subset selection strategies are also shown.

The modality that individually gave the best classification performance was physiology, with a classification rate of 75.94%. Audio was the worst, with 39.07%. Furthermore, the modality combinations giving the highest classification performance were physiology + video and physiology + video + audio, with a classification rate of 79.30% for both combinations. These results were all obtained using the full set of features within the specified modalities. When taking the subset selection methods into account, the best result was obtained using the feature subset resulting from the best-first search subset selection strategy, giving a classification rate of 82.24%. The selected features were 10 physiology features, 19 video features and no audio features. These features are listed in Table 2.3.
Figure 2.7: Classification results of CVO method using the SVM classifier. The results are sorted in ascending order. \( p + v + a \) stands for physiology + video + audio.

Table 2.3: The resulting features using best-first search for subset selection.

<table>
<thead>
<tr>
<th>Physiology</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSR_SCL_mean</td>
<td>Confidence</td>
</tr>
<tr>
<td>GSR_SCR_NR_per_sec</td>
<td>Face Standard Deviation</td>
</tr>
<tr>
<td>Temp_mean</td>
<td>Right Eye Ratio</td>
</tr>
<tr>
<td>Temp_std</td>
<td>Mouth Ratio</td>
</tr>
<tr>
<td>Temp_slope</td>
<td>Face Yaw</td>
</tr>
<tr>
<td>ECG_IBI_mean</td>
<td>Right Pupil Y</td>
</tr>
<tr>
<td>ECG_IBI_pow_LF</td>
<td>Left Pupil Y</td>
</tr>
<tr>
<td>ECG_IBI_pow_HF</td>
<td>Lower Lip Center Y</td>
</tr>
<tr>
<td>ECG_IBI_RMSSD</td>
<td>Right Mouth Corner X</td>
</tr>
<tr>
<td>ECG_IBI_SDNN</td>
<td>Right Mouth Corner Y</td>
</tr>
<tr>
<td></td>
<td>Left Mouth Corner X</td>
</tr>
<tr>
<td></td>
<td>Left Mouth Corner Y</td>
</tr>
<tr>
<td></td>
<td>Left Outer Eye Corner Y</td>
</tr>
<tr>
<td></td>
<td>Right Outer Eye Corner X</td>
</tr>
<tr>
<td></td>
<td>Left Lower Lip Y</td>
</tr>
<tr>
<td></td>
<td>Gaze Tilt</td>
</tr>
<tr>
<td></td>
<td>Gaze Pan</td>
</tr>
<tr>
<td></td>
<td>Left Eye Openness Level</td>
</tr>
<tr>
<td></td>
<td>Mouth Openness Level</td>
</tr>
</tbody>
</table>

The confusion matrix in Table 2.4 shows the classification results of best-first search per class. Sad was classified most accurately, followed by Angry. Neutral had the worst classification result.
Table 2.4: Confusion matrix of 5-fold cross-validation result with the SVM classifier, using best-first search feature subset.

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Happy</th>
<th>Relaxed</th>
<th>Sad</th>
<th>Angry</th>
<th>Neutral</th>
<th># Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>78.52%</td>
<td>6.67%</td>
<td>6.67%</td>
<td>5.93%</td>
<td>2.22%</td>
<td>135</td>
</tr>
<tr>
<td>Relaxed</td>
<td>7.27%</td>
<td>79.09%</td>
<td>5.45%</td>
<td>7.27%</td>
<td>0.91%</td>
<td>110</td>
</tr>
<tr>
<td>Sad</td>
<td>1.68%</td>
<td>1.68%</td>
<td>88.83%</td>
<td>6.15%</td>
<td>1.68%</td>
<td>179</td>
</tr>
<tr>
<td>Angry</td>
<td>4.10%</td>
<td>2.05%</td>
<td>4.10%</td>
<td>86.67%</td>
<td>3.08%</td>
<td>195</td>
</tr>
<tr>
<td>Neutral</td>
<td>9.38%</td>
<td>2.08%</td>
<td>8.33%</td>
<td>10.42%</td>
<td>69.79%</td>
<td>96</td>
</tr>
</tbody>
</table>

Physiology sub-modalities

As can be observed from Table 2.2, physiology was the modality giving the best individual classification performance. Since physiology is a broad modality, it can be considered as consisting of a set of 'sub-modalities': Skin Conductance (SC), Temperature (Temp), Respiration (RSP) and Electrocardiogram (ECG). The number of features representing each sub-modality are SC: 4, Temp: 3, RSP: 1 and ECG: 6. The added value of each of these sub-modalities was investigated, by testing the performance when removing the features of each of these sub-modalities from the total stack of physiology features. The results are shown in Table 2.5.

Table 2.5: Classification results of 5-fold cross-validation with the SVM classifier for different physiology sub-modalities.

<table>
<thead>
<tr>
<th>Physiology sub-modalities</th>
<th>SC</th>
<th>Temp</th>
<th>RSP</th>
<th>ECG</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>59.58%</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td>77.62%</td>
</tr>
<tr>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td>65.73%</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>67.41%</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x</td>
<td>75.94%</td>
</tr>
</tbody>
</table>

The results suggest ECG to be the strongest sub-modality, followed by Temp and SC. RSP appears to be the worst sub-modality — and, at the same time the worst feature — since the classification performance increased when this feature was removed, whereas feature removal of any of the other modalities resulted in a decrease in classification performance.

Valence/arousal relabeled classes

Using the first method for re-labeling the emotion classes resulted in a classification performance of 79.30%, which is exactly equal to that of using the original 5 emotion classes. The confusion matrix is shown in Table 2.6. Angry was classified best [92.86%], Sad was classified worst [46.43%]. Sad was most often classified as Angry.

In the second re-labeling method, the emotion labels were replaced by valence labels (positive/negative), using subjects' overall valence ratings. 337 feature vectors were labeled positive and 378 were labeled negative. When using only physiology features, classification rate was 82.80%. When using only video features performance was 76.22%. Finally, in the third
Table 2.6: Confusion matrix of SVM classification result using re-computed emotion classes based on the overall valence- and arousal ratings.

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Happy</th>
<th>Angry</th>
<th>Relaxed</th>
<th>Sad</th>
<th>#Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>74.21%</td>
<td>22.17%</td>
<td>3.17%</td>
<td>0.45%</td>
<td>221</td>
</tr>
<tr>
<td>Angry</td>
<td>6.29%</td>
<td>92.86%</td>
<td>0.86%</td>
<td>0.00%</td>
<td>350</td>
</tr>
<tr>
<td>Relaxed</td>
<td>18.10%</td>
<td>25.86%</td>
<td>56.03%</td>
<td>0.00%</td>
<td>116</td>
</tr>
<tr>
<td>Sad</td>
<td>10.71%</td>
<td>35.71%</td>
<td>7.14%</td>
<td>46.43%</td>
<td>28</td>
</tr>
</tbody>
</table>

re-labeling method, the emotion labels were replaced by arousal labels (active/passive) using subjects’ overall arousal ratings. In total, 571 feature vectors were labeled active and 144 were labeled passive. When using only physiology features, classification rate was 85.87%. When using only video features performance was 78.60%.

2.3.2 MLP Results

In Table 2.7 the classification results are shown for the Multi-Layer Perceptron. The modality that individually gave the best classification performance was video [53.63%], followed by physiology [53.43%]. Audio was again the modality giving the lowest classification performance [36.98%]. A visualization of the results is shown in Figure 2.8. The best classification performance was obtained with the physiology + video combination (i.e. the combination of all physiology- and all video features), giving a classification rate of 73.71%.

Table 2.7: An overview of the classification results for the MLP classifier for every modality combination

<table>
<thead>
<tr>
<th>NO FEATURE SELECTION</th>
<th>Physio.</th>
<th>Video</th>
<th>Audio</th>
<th>CVO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td>53.43%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x</td>
<td></td>
<td>53.63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>36.98%</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td>73.71%</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td>55.94%</td>
</tr>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>71.33%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENETIC SEARCH</th>
<th>Physio.</th>
<th>Video</th>
<th>Audio</th>
<th>CVO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>72.73%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BEST FIRST SEARCH</th>
<th>Physio.</th>
<th>Video</th>
<th>Audio</th>
<th>CVO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>70.21%</td>
</tr>
</tbody>
</table>
Figure 2.8: Classification results of CVO method with the MLP classifier. The results are sorted in ascending order. \( p + v + a \) stands for physiology + video + audio.

Table 2.8 shows the confusion matrix of the classification result for the physiology + video combination. As was the case for the SVM classifier, Sad here also has the best prediction score. Relaxed was classified worst.

Table 2.8: Confusion matrix of of 5-fold cross-validation results with the MLP classifier, using the physiology + video combination.

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Happy</th>
<th>Relaxed</th>
<th>Sad</th>
<th>Angry</th>
<th>Neutral</th>
<th># Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>69.63%</td>
<td>5.93%</td>
<td>8.15%</td>
<td>8.89%</td>
<td>7.41%</td>
<td>135</td>
</tr>
<tr>
<td>Relaxed</td>
<td>13.64%</td>
<td>60.91%</td>
<td>6.36%</td>
<td>16.36%</td>
<td>2.73%</td>
<td>110</td>
</tr>
<tr>
<td>Sad</td>
<td>4.47%</td>
<td>3.35%</td>
<td>82.68%</td>
<td>7.82%</td>
<td>1.68%</td>
<td>179</td>
</tr>
<tr>
<td>Angry</td>
<td>6.67%</td>
<td>4.10%</td>
<td>4.62%</td>
<td>78.97%</td>
<td>5.64%</td>
<td>195</td>
</tr>
<tr>
<td>Neutral</td>
<td>6.25%</td>
<td>4.17%</td>
<td>12.50%</td>
<td>10.42%</td>
<td>66.67%</td>
<td>96</td>
</tr>
</tbody>
</table>

Valence/arousal relabeled classes

Using the first method for re-labeling the emotion classes resulted in a classification performance of 74.55%. The confusion matrix is shown in Table 2.9. Relabeling the emotion labels to valence labels (337 positive and 378 negative) resulted in a classification performance of 72.45% when using only physiology features, and 72.73% when using only video features. Finally, relabeling the emotion classes to arousal labels (571 active and 144 passive) resulted in classification performances of 80.70% and 77.90% for physiology -and video features, respectively.
Table 2.9: Confusion matrix of MLP classification result using re-computed emotion classes based on the overall valence- and arousal ratings.

<table>
<thead>
<tr>
<th>Predicted class</th>
<th>Happy</th>
<th>Angry</th>
<th>Relaxed</th>
<th>Sad</th>
<th>#Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual class</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anger</td>
<td>0.57%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relaxed</td>
<td>10.10%</td>
<td>45.52%</td>
<td></td>
<td>5.29%</td>
<td>116</td>
</tr>
<tr>
<td>Sad</td>
<td>0.71%</td>
<td>17.86%</td>
<td>14.29%</td>
<td></td>
<td>28</td>
</tr>
</tbody>
</table>

2.4 Discussion

In this chapter we presented our work on multi-modal emotion recognition. First, data was gathered from subjects telling about an emotional event from their life, for five emotions: happy, relaxed, sad, angry and neutral. The modalities from which emotion data was gathered were video (for capturing facial expressions), audio (for capturing vocal expressions) and physiology (Appendix A). Features were extracted for each modality separately. For extracting facial features, a software tool (OKAO Vision) was used for tracking and extracting facial feature points, as well as position information of the face. For extracting audio features, the software tool PRAAT was used. The physiological signals from which features were extracted were skin conductance (SC), temperature (Temp), respiration (RSP), and electrocardiogram (ECG). The physiological features were extracted using a MATLAB implementation developed at Philips Research Eindhoven. All features were extracted over a time window of 10 seconds, with the exception of the ECG features, for these a 30 second time window was required in order to get reliable feature values.

The features from the three modalities were combined using a fusing method called feature-level fusion, meaning that the features from different modalities were concatenated. In order to find the most useful features, and to filter out the ones that cause noise, feature selection was applied. Two selection methods were used. The first was a wrapper approach for finding the best feature subset. Two search algorithms were used for this, best-first search and genetic search. The other selection method consisted of making a ranked list of the entire feature set using the Chi-square statistic, and evaluating different subsets of this list. After feature selection, classification was performed, using two classifiers, a Support-Vector Machine (SVM) and a Multi-Layer Perceptron (MLP).

Three classification methods were applied, Leave-one-Out (LVO): training on the data of all but one subject and testing on the left-out subject; Cross Validation per Subject (CVS): 5-fold cross validation applied to the data of all subjects separately; and Cross Validation Overall (CVO): 5-fold cross validation applied to the combined data of all subjects. For the MLP classifier, only the latter method was applied. The results showed that, when looking at each modality separately, physiology gave the best results. This was most prominent for the SVM classifier. The best modality combination appeared to be the combination of video and physiology.

We also investigated the claim by Wilder (1962) that arousal (how active an emotion is) is best predicted by physiology whereas valence (how positive or negative an emotion is) is best predicted by facial expressions. The results by Wilder (1962) could not be reproduced. The results are more elaborately discussed below.
Classification methods

For the SVM several observations can be made. When comparing the results per classification method, we can see that CVS gave the best results. Great interpersonal differences in the data, even after scaling, cause much harder generalizability over a larger group of persons. This brings us to the point of timing effects. In order to ensure that the differences in the data for different emotional stories were caused by the experienced emotions — instead of by the moment in time when the stories were told — the order of the emotions for which the subjects were instructed to tell a story, was counterbalanced. As a consequence, the differences in the data for different emotions, taken over multiple subjects, should be caused by the emotions themselves. However, timing effects within subjects may still have been present, which could have emphasized the differences between the data of different emotions. This may have caused the relatively high classification results for CVS. In order to check the effect of time versus the effect of emotions on the resulting data, a possible approach would have been to let each person tell multiple stories for the same emotion at different moments in time.

The generalizability problem also becomes clear when looking at the results for the LVO method, which were clearly worse than those for CVS and CVO. Apart from the fact that in the CVO method cross validation was used, whereas in the LVO one training set and one test set was used, the only difference between these methods is that in CVO, the data from the test subject is also present in the training set, whereas in the LVO it is not. In order to narrow the performance gap between LVO and CVO (i.e. in order to improve generalizability) a (much) larger subject group would be needed, since this could cause the data coming from different subjects to become more alike.

Modalities used

Data gathered via three modalities was used: video, audio and physiology. When comparing the differences in performance between these modalities, the most important observations to be made are that: for the SVM the performance when using only (and at the same time, all) physiology features was considerably higher than when only video or audio features were used; for both SVM and MLP the audio features gave relatively bad results; and the best classification result was obtained when using a combination of physiology and video features. The relatively low classification performances for audio could be explained by the fact that, when telling about an event when some emotion was experienced, one may feel constrained in using the same vocal expressiveness as one does in the actual event (in case of a high arousal emotion, such as happy and angry) since (1) loudly expressing oneself could draw much unwanted attention from the environment and (2) the subjects from which the data was gathered were not talking to a real person and therefore a strong incentive for vocal expressiveness was missing. The first point seems to be the most plausible one, since the second point would imply that the classification performance for video and audio would be approximately equal. Moreover, the first point would explain the relatively good performance for the physiology features, since, according to K. Kim et al. (2004) and J. Kim and André (2008), physiological processes cannot be deliberately controlled as easily as facial and vocal expressions, and therefore, they are harder to constrain.

Valence/arousal relabeled classes

The results of Wilder (1962), showing that valence could best be predicted by facial expressions and that physiological signals would be a better indicator for arousal, could not be reproduced. For the SVM classifier, physiological features gave the best results for both the valence and the arousal classes. For the MLP classifier, the result difference between using video and physiolog-
ical features for the valence classes was too small to draw any conclusions from it, and for the arousal relabeled classes, the performance for the physiological features was only slightly better than the performance for the video features.

In sum, we have seen that combining information from multiple modalities can considerably increase overall classification performance. Furthermore, for automated emotion recognition, the use of physiological features appeared to greatly increase classification performance, especially for the SVM classifier. However, when judging someone’s emotional state, humans use mainly facial and vocal information (Mehrabian, 1968). Hence, the question could be asked whether the perception of a physiological signal could influence this judgment. This question is the starting point for the next chapter, in which our research on the influence of a physiological feature on the experience of emotional intensity is described.
Chapter 3

Emotion Intensity Judgment

3.1 Introduction

In real-life situations, when two modalities send contradictory signals, the message sent by one modality could undo the message sent by the other. This may be the case when one tries to be ironic (or “funny”), for example, when one is looking happy, while sounding serious. When we do this, we generally assume that the person receiving these signals understands the contradiction, and knows how to interpret the message. However, in this example physiological signals are not taken into the equation. When people are nervous (e.g. during a presentation), they can try to hide this by keeping their facial and vocal expressiveness under control, so that they appear calm. Nevertheless, their physiology may tell otherwise. Several researchers (J. Kim & André, 2008; K. Kim et al., 2004) agree about the fact that physiological signals, which represent the activity in the autonomic nervous system (ANS), more directly reflect an emotional state than, for example, facial or vocal expressions, and they are therefore harder to deliberately control. One could wonder how contradictions between appearance (e.g. facial expressions) and physiological signals are interpreted when it comes to judgment of emotional intensity. This is in fact one of the questions we try to answer in this research project.

In the previous chapter, a research is explained that focused on automated emotion classification using video, audio and physiological features as input. It was shown that combining information from multiple modalities improved overall classification performance, and that especially the use of physiological features resulted in a performance improvement. The experiment described in this chapter is about emotion judgment by humans. For judging a persons emotions, we mainly pay attention to this person’s facial expressions, and voice. But, as noted by several researchers (Picard et al., 2001; Sebe et al., 2005), physiological signals, like heart rate, skin conductance and respiration rate can also be used for emotion recognition in human-human interaction. However, for most of these signals to be observed one has to be very close to, or even touch, the person from which they are observed, as opposed to facial or vocal signals, which can be read from a much greater distance. Since the influence of physiological signals is relatively small, and since these signals had such a great influence on automated emotion recognition, the question is investigated what the influence of physiological signals would be if they would be amplified, so that their perception would be reinforced.

In the experiment described in this chapter, participants are given the task to judge the intensity of another persons experienced emotion, based on the persons facial expressions and an added physiological signal. The physiological signal that is selected for this is a heartbeat. Of all physiological signals (skin conductance (SC), skin temperature (Temp), respiration (RSP),
heart rate (HR), etc.), HR is probably the most well known to people for indicating emotional intensity (Kreibig, 2010).

The main research goal of the experiment presented in this chapter is to find out to what extent a physiological signal would influence peoples judgment of someone else’s emotional state, in case they would be able to perceive this signal more easily, e.g. by presenting the physiological signal in auditory form. Another research goal is to find out to what extent people trust a person’s physiological signal, compared to his/her facial expressions, in case these two contradict. Furthermore, the experience of the amplification of the physiological signal — which in normal situations can not be perceived that easily — is also investigated. The next subsection describes a set of research questions that arose from these research goals.

3.1.1 Research questions

There are several research questions we try to answer in this chapter. These are divided into a Main research questions and several Secondary research questions. These are as follows:

Main research question:
Can the judgment of someone’s emotional intensity significantly be influenced by a physiological signal (e.g. a heartbeat)?

Secondary research questions:

1. What is the influence of the physiological signal in case it is in agreement with the facial information? (e.g. fast heart rate and angry facial expressions)
2. What is the influence of the added physiological signal in case it contradicts the facial information? (e.g. fast heart rate and calm facial expressions)
   (a) Which information will people use more, facial expressions or physiology?
   (b) Will people use the information of one modality (either facial expression information or physiology) or will they use a combination (e.g. a rough average)?
3. How do people experience the physiological signal being added? (i.e. will they find it useful, irrelevant or interfering for/with their assessment of emotional intensity?)

In the following, our hypotheses for each of these research questions are described.

3.1.2 Hypotheses

There are reasons to believe that facial expressions will have more influence on people’s judgment of a person’s emotional intensity than a physiological signal (like heartbeat) would, and vice versa. Some reasons in favor of facial expressions and some reasons in favor of heartbeat are described below. It is important to note that the experiment described in this chapter is not about guessing the displayed emotion, but about judging the emotional intensity of the person displaying the emotion. Therefore it may not be self-evident to assume the same influence of physiology in this experiment as in the one described in the previous chapter.

Some arguments in favor of facial expressions are the following:

- Judging a person’s affective state is mainly done through observing and interpreting facial and vocal expressions (Mehrabian, 1968). The contribution of physiology is relatively small.
For this reason participants may be inclined to use the facial expressions as their primary cue, and ignore the physiological signal.

- Facial expressions can quite easily be associated to specific emotions, at least in case of the 6 basic emotions — happiness, sadness, anger, fear, disgust and surprise — proposed by Ekman (1971), whereas for physiological signals this is not always the case. A fast heartbeat for example, can be associated to both positive and negative affective states.

Some arguments in favor of a physiological signal are:

- People are able to deliberately change their facial expression to one that does not correspond to their currently felt emotion, (some people can do this better than others). Therefore, people will probably not unthinkingly trust facial expressions that are being displayed by other people. They will also look at cues like body posture and body movement, in order to see if these correspond to the displayed facial expression. However, these can also be controlled deliberately. According to several researchers (Healey, 2002; J. Kim & André, 2008; K. Kim et al., 2004) physiological signals are not that easy to change at will. Therefore, people may have more trust in a physiological signal when it comes to judging emotional state.

- Facial expressions show specific emotions, and whether these are positive or negative (i.e. the valence of an emotion). Variations in physiological signals mainly indicate arousal of an emotion. Since this experiment is about judging emotional intensity, arousal may therefore be the most reliable measure to use and so people may see the physiological signal as the most informative cue.

From these arguments it appears that no clear-cut decision can be made about which modality would be most influential when it comes to judging emotional intensity. But given the arguments for both modalities we can now try to state our hypotheses for the above described research questions.

**Main research question:** *Can the judgment of someone’s emotional intensity significantly be influenced by a physiological signal?*

This research question is only about the influence of physiology. From the arguments in favor of physiology one could assume that a physiological signal would definitely influence people's judgment. Since physiological signals are hard to be manipulated at will, people will probably trust the credibility of these signals, when it comes to representing a person's affective state. Furthermore, the association between the selected physiological signal — a heartbeat — and emotional intensity will probably be self-evident to most people. And so, when it comes to the main research question, our hypothesis is that a physiological signal can significantly influence people's judgment of the emotional intensity.

**Secondary research questions:**

1. **What is the influence of the added physiological signal in case it is in agreement with the facial information?**

   In case facial expressions and physiology (heart rate) of a person are in agreement, there are two possible combinations:

   - little facial expressions + slow heart rate
• strong facial expressions + fast heart rate

Since in this condition both modalities are in agreement, there is no real reason to assume that the person from which they (presumably) originate is hiding his/her true emotional state. In other words, if both physiology and facial expressions of a person indicate calmness, then the person is probably feeling calm. Likewise, if physiology and facial expressions indicate high tension, then the person is probably feeling tense. We can only imagine that participants will apply this reasoning strategy. Hence, our hypothesis for this research question is that in case of strong facial expressions + fast heart rate the felt emotion will be judged significantly more intense than in case of little facial expressions + slow heart rate.

2. **What is the influence of the added physiological signal in case it contradicts the facial information?**

(a) Which information will people use more, facial expressions or physiology?

(b) Will people use the information of one modality (either facial expression information or physiology) or will they use a combination (e.g. a rough average)?

In case facial expressions and heart rate of a person are contradictory, there are two possible combinations:

- little facial expressions + fast heart rate
- strong facial expressions + slow heart rate

From the arguments in favor of facial expressions and from the arguments in favor of physiology it is hard to say which modality would be most influential in case the modalities would contradict each other. Another uncertainty is how people will combine the information from the conflicting signals. There is the possibility that people will choose to use the information of one modality, but there is also the possibility that people will combine the information of both modalities by taking a rough average. As explained above, people are capable of deliberately controlling, and therefore able to lie with, their facial expressions whereas for physiological signals this is not quite as easy. Thus, in case of a contradiction, people may feel inclined to trust the physiological signal more than the facial expressions. And so, our hypothesis is that the combination of little facial expressions + fast heart rate will be judged significantly more intense than strong facial expressions + slow heart rate.

3. **How do people experience the physiological signal being added?**

Hearing the heartbeat of the person in front of you is an experience that is probably new to most people and may therefore, at first, be felt as strange and uneasy. After a short while however, people may get used to it, since, although not in this setting, the sound of a heartbeat is not new to people. In movies, for example, it is sometimes used to increase tension during a thrilling scene. That is why, after a short familiarization period, the feeling of strangeness will probably disappear, and people will be able to take into account the heartbeat signal for their emotional intensity judgment. Our hypothesis is therefore that the physiological signal will be experienced as helpful.
3.2 Methods

3.2.1 Participants

Participants were 42 students, 10 male and 32 female, aged 18 to 30 years. For their participation they received course credit or 5 euros. Before the start of the experiment, subjects signed an informed consent which stated that their participation was voluntary and that they could quit the experiment at any time without any negative consequences concerning their reward. At the end of the experiment, which is further described in the following sections, the participants were given two open ended questions — in the form of a questionnaire — about the contents of the experiment. The questions are described in Section 3.2.4. Based on the given answers, seven participants were removed, due to the suspicion that they might have realized the true goal of the experiment.

3.2.2 Experimental design

In the experiment, participants were looking at videos showing persons telling about an emotional event from their life. Each video was accompanied by a heartbeat sound. There were two conditions for the videos: either it was a person telling about a neutral event (e.g. going home) or it was a person telling about an event when (s)he was angry. For the heartbeat sounds — which were played during the videos — there were also two conditions, slow [67.5 – 72 bpm] and fast [80 – 84.5 bpm]. The experiment consisted of 10 parts, each in which two videos were shown of two different persons, each accompanied by a different heartbeat sound. The combination of facial expressions (neutral/angry) and heartbeats (slow/fast) resulted in 4 possible conditions for each video. With these, (4 x 4 =) 16 pair combinations were created. By means of counterbalancing, 10 of these 16 video pair combinations were selected for each participant. In total, each participant saw 20 videos, coming from 10 different persons. The two videos from one person never occurred in the same pair.

3.2.3 Materials

Videos

The videos were a selection of those used for the investigation on Multi-modal Emotion Recognition, described in Chapter 2. For details about these videos, we refer the reader to Appendix A. The emotion displayed in a video was either a high arousal emotion or no emotion (neutral). The reason for comparing these two conditions is that they are the most clearly separable by both video (almost no facial expressiveness vs strong facial expressiveness), and heartbeat (slow vs fast). For the high arousal emotion, videos of persons telling about an angry event were selected, since the arousal ratings were highest for this emotion. The selection of the videos was made based on the self-rated arousal values of the persons in the videos. The videos of the persons with the greatest absolute arousal difference between the neutral and the angry video were selected. In total, the neutral and angry videos of 10 persons were selected. Of those videos, the 30 seconds where the arousal ratings were the lowest (for neutral) or the highest (for angry) were taken to be used for the experiment. However, since the videos were shown in pairs, the duration of the videos was shortened to 10 seconds in order to prevent the memory of the first shown video to fade away when watching the second video. The resolution of the videos was 640 x 480 pixels.
**Heartbeats**

The heartbeat displayed during the video was either slow [67.5 – 72 bpm] or fast [80 – 84.5 bpm], and remained constant during the entire video. For each condition, 10 heartbeat files were created, with a different heart rate for each file. The lowest heart rate matched the lower bound of the indicated range (67.5 for slow and 80 for fast), and was increased with 0.5 bpm per file, until the upper bound of the range was reached. All heartbeat files were created from one file containing an artificial heartbeat with a rate of approximately 70 bpm.

**Double Visual Analogue Scale**

Using a double Visual Analogue Scale (VAS), participants were able to rate the emotional intensity of the persons they saw in the videos. The double VAS consists of two continuous scales, ranging from 1 (Very low emotional intensity) to 9 (Very high emotional intensity), which were placed next to each other. On these scales, participants could give their emotional intensity rating for both videos in the just seen video pair. The double VAS that was used in this experiment is shown in Figure 3.1.

![Double Visual Analogue Scale](image)

**Empathy questionnaire**

A modified version of the Multi-Dimensional Emotional Empathy Scale (Caruso & Mayer, 1998) was used for measuring empathy, in which 30 statements could be answered on a 9-point scale,
instead of the original 5-points scale. The modified scale went from -4 (very strong disagreement) to 4 (very strong agreement). The questionnaire is shown in Appendix D. The goal of using this questionnaire is to make a distinction between ratings from persons with a high empathy score compared to persons with a low empathy score.

**Hardware**

Participants were sitting in front of a 17 inch LCD screen displaying the slides for the experiment. It was equipped with a Logitech Premium Stereo headset, a mouse and a keyboard. The headset was used for presenting the heartbeat. The sound volume of the headset was kept the same for all participants. The mouse could be used for giving the emotional intensity rating on the double VAS, and to move on to the next slide, by pressing a Continue button on the screen. The keyboard was for answering open ended questions.

**3.2.4 Procedure**

Participants signed an informed consent form, after which they were given instructions about the experiment and they had the opportunity to ask questions if anything was unclear. The instruction form is shown in Appendix C. The participants were told that the goal of the experiment was to “investigate how people judge emotional intensity”. On the computer screen the instructions for the experiment were also displayed. In the instructions the participants were explained that the experiment consisted of 10 parts, each in which they would see 2 videos of 2 different persons telling about an emotional event from their life. The pair of videos would be repeated 3 times. The participants were informed that audio would be turned off and participants would hear the heartbeat of the person in the video. Finally, they would be asked to give a judgment of the intensity of the emotion that was felt by both persons, and that they would have to do so on a scale, ranging from 1 to 9, using slidebars. After reading the instructions from the screen, the participants were asked whether they had any questions. Then, before actual start of the experiment, a practice session was started, in which participant could get familiar with the video + heartbeat combination. The practice sessions, analogous to the parts in the actual experiment, consisted of a pair of videos — each accompanied by a heartbeat sound — which was repeated 3 times, followed by a double Visual Analogue Scale (VAS) (Figure 3.1). The videos that were shown in the practice session were the same for all participants. These videos were not used in the actual experiment. During the practice session, the experimenter stayed with the participant to give him/her the opportunity to ask questions in case something was still not clear. After the practice session, the experimenter left the room and the experiment started. After participants had seen and rated all 10 video pairs, they were given the Multi-Dimensional Emotional Empathy Scale questionnaire, followed by two open questions. The first open question asked the participants what (s)he thought the experiment was about. The second

![Figure 3.2: Experimental procedure.](image-url)
question asked the participants how (s)he felt about hearing the heartbeats of the persons in the videos. Answering this question concluded the experiment. A visualization of the experimental procedure is shown in Figure 3.2. After the experiment, participants were debriefed. They were told that the heartbeats they heard were not really coming from the person in the videos, and they were explained what the real purpose of the experiment was. The entire experiment took about 25 minutes per participant.

3.3 Results

Within-subject effects

Using a repeated measures analysis of variance (ANOVA), the effects of facial expressions (neutral vs angry) and heart rate (slow vs fast) on emotional intensity were investigated. Angry facial expressions resulted in significantly higher intensity ratings than neutral facial expressions ($F = 35.93, p < 0.05, \eta^2 = 0.51$). The effect of heart rate was also significant ($F = 9.60, p < 0.05, \eta^2 = 0.22$), meaning that higher heart rates caused higher emotional intensity ratings than lower heart rates. There was no interaction effect ($F = 0.122, p = 0.73, \eta^2 = 0.00$). The main results of the experiment are shown in Figure 3.3, where the average intensity rating per facial expression + heart rate combination is presented.

![Figure 3.3: Emotional intensity ratings per facial expression + heart rate combination.](image)

Table 3.1 shows the mean, standard deviation, minimum and maximum of the average ratings per participant, for each of the four conditions. For example, the Rating minimum/maximum depict the minimum/maximum average value that a subject had given for this combination. From this table it can be seen that the combination of neutral facial expressions and fast heart rate led to the largest diversity of emotional intensity ratings. For this combination the standard deviation ($sd = 1.23$) as well as the absolute difference between the maximum and minimum rating ($diff = 5.38$) is highest. This is followed by the condition Neutral + Slow, with a standard deviation of 1.16 and a max-min difference of 5.22. The condition Angry + Slow gave the lowest
standard deviation value (1.07) and a max-min difference of 4.2. Finally, the condition Angry + Fast resulted in the lowest max-min difference (3.96), and a standard deviation of 1.08.

Table 3.1: The results of the emotional intensity judgment experiment, showing the mean, standard deviation, minimum and maximum of the average rating values per subject for each facial expression + heart rate combination.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Heart rate</th>
<th>Rating mean</th>
<th>Rating std. dev.</th>
<th>Rating minimum</th>
<th>Rating maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td>Slow</td>
<td>4.64</td>
<td>1.16</td>
<td>1.76</td>
<td>6.98</td>
</tr>
<tr>
<td>Neutral</td>
<td>Fast</td>
<td>5.17</td>
<td>1.23</td>
<td>2.21</td>
<td>7.59</td>
</tr>
<tr>
<td>Angry</td>
<td>Slow</td>
<td>5.25</td>
<td>1.07</td>
<td>3</td>
<td>7.2</td>
</tr>
<tr>
<td>Angry</td>
<td>Fast</td>
<td>5.87</td>
<td>1.08</td>
<td>3.98</td>
<td>7.93</td>
</tr>
</tbody>
</table>

Effect of empathy

As explained in the previous section, participants were given a questionnaire – the Multi-Dimensional Emotional Empathy Scale questionnaire – at the end of the experiment to measure their empathy. The results of this questionnaire have been used to see whether there was any correlation between their overall empathy score and their rating behavior based on differences in facial expressions and/or heart rates. More specifically, the goal was to see whether persons with higher empathy values showed greater differences between their emotional intensity ratings for angry and neutral facial expressions and/or between their emotional intensity ratings for fast and slow heart rates. In Figures 3.4 and 3.5, scatter plots show the average difference scores of emotional intensity ratings, for angry and neutral facial expressions (Figure 3.4), and for fast and slow heart rates (Figure 3.5), for every subject, plotted against their emotional empathy score.

A regression analysis was performed to see whether the difference between emotional intensity ratings for angry and neutral facial expressions could be predicted by one’s emotional empathy score. It appeared that the proportion of variability of the differences between intensity scores for angry and neutral facial expressions, as a result of differences in empathy ($R^2 = 0.083$) was marginally significantly greater than zero ($F(1,33) = 3.005, \beta = 0.289, p < 0.1$). Furthermore, a regression analysis was performed to see whether the difference between emotional intensity ratings for fast and slow heart rates could be predicted by one’s emotional empathy score. The proportion of variability of the differences between intensity scores for fast and slow heart rates, as a result of differences in empathy ($R^2 = 0.026$) was not significantly greater than zero ($F(1,33) = 0.870, \beta = 0.160, p = 0.36$).

Heartbeat experience

To the questionnaire asking the participants how they felt about hearing the heartbeat of the persons in the videos, answers varied greatly from person to person. A number of participants indicated that they did not like hearing the heartbeat (e.g. because it made them nervous), while others experienced it as pleasant and relaxing. Some indicated that hearing the heartbeats helped them to make their judgment, whereas to others the heartbeats were confusing, and did not always seem to agree with the emotion expressed in the video. There were also participants who reported having no special feeling about hearing the heartbeats. In sum, the results showed that answers to this questionnaire varied greatly from person to person.
Figure 3.4: Scatter plot showing the average difference between the emotional intensity scores for angry and neutral facial expressions for every subject, plotted against their emotional empathy score.

Figure 3.5: Scatter plot showing the average difference between the emotional intensity scores for fast and slow heart rates for every subject, plotted against their emotional empathy score.

### 3.4 Discussion

In this chapter we described an experiment in which people judged another person’s emotional intensity based on their facial expressions and heart rate. The experiment showed promising
results, since both facial expressions, (and perhaps more importantly) heart rate showed a signific.
ificant effect on emotional intensity judgment. we will now first answer the research questions that were presented in the Introduction of this chapter. Then we will give our view on the potential of physiology for human-human communication. We will conclude this chapter by giving some future recommendations.

3.4.1 Influence of physiological signal

In the Introduction, we presented four research questions. With the results presented in the previous section, these questions can now be answered. We will first answer the secondary research questions in the order they were presented in the Introduction, and conclude by answering the main research question.

What is the influence of the physiological signal in case it is in agreement with the facial information?

In case the facial expressions and heart rate were in agreement, they appeared to strengthen one another’s effect. When neutral facial expressions were shown (i.e. a video of a person telling about a neutral event), combined with a slow heart rate sound, the intensity ratings were significantly lower than in the other conditions. Angry facial expressions combined with a fast heart rate resulted in the highest emotional intensity ratings.

An important side note that should be made here is that for the condition consisting of neutral facial expressions + slow heart rate the standard deviation of the emotional intensity rating was relatively high. Possibly, to some, the presence of a heartbeat sound resulted in a feeling of high tension, regardless of its rate.

What is the influence of the added physiological signal in case it contradicts the facial information?

When the information from the two channels contradicted, no factor appeared to have the greatest effect. From the results it appeared that, generally, people did not choose to go for the information from one modality while ignoring the other, but seemed to have taken a middle way in order to concede to the information from both modalities. From the results it can be seen that there is no clear preference, when looking at which modality people trust more. The difference in intensity rating for neutral facial expressions + fast heart rate and angry facial expressions + slow heart rate is too small to draw any conclusions from it.

However, the condition neutral facial expressions + fast heart rate resulted in the highest standard deviation value, indicating that the ratings for this condition varied the most among participants. This is interesting, since this could imply that, to some participants, it may have appeared that the person in the video was trying to hide some tense emotional state (e.g. nervousness), whereas other participants may not have applied this method of reasoning.

How do people experience the physiological signal being added?

It appears that no clear-cut answer can be given to this research question, since answers are too diverse. It seems that the way a physiological signal (in this case, a heartbeat) is experienced, depends on the person experiencing it. The experience may also vary per condition. It is not unlikely that, in conditions where facial expressions and heart rate contradict, the heartbeat sound may be experienced as more disturbing/interfering than in conditions where the modalities are
in agreement. However, without any hard evidence to back this up, the latter is only a guess.

Now that we have answered all secondary research questions, we have come to the point to an-
swer the main research question:

*Can the judgment of someone’s emotional intensity significantly be influenced by a physiological
signal?*

From the results presented in the previous section we can clearly say that the answer to this
question is yes. The emotional intensity ratings for the facial expressions accompanied by fast
heartbeats were significantly higher than those accompanied by slow heartbeats.

When looking back at the hypotheses presented in Section 3.1.2, it appears that the results of
the experiment were not always as expected. In case the heartbeat sound contradicted with
the facial expressions, the participants did not seem to trust only one (the heartbeat) as we
hypothesized, but instead, they used the information of both modalities and selected a middle
way to more or less concede with both. Also, the experience of the heartbeat sound was not
as expected. We hypothesized that, after a short while, people would get used to it and would
definitely not feel strange anymore. Yet, there was still a number of participants didn’t appear
to like the experience of the hearing the heartbeat sound.

### 3.4.2 Physiology in human-human interaction

The results show high potential for physiological signals to be used for communicating emotions.
We have seen that people generally tend to adjust their judgment of someone’s emotional state
using information from a physiological signal (in case this signal was presented in form of a
heartbeat sound). The participants of the experiment appeared to know how to interpret the
physiological signal and were able to adjust their intensity judgment based on this signal. Nu-
merous applications can be thought of which could benefit from the integration of a physiological
into the human-human communication process. Examples are applications for instant messaging,
where next to text, video and audio, the addition of a physiological signal could give you a better
understanding of the affective state of your communication partner(s).

However, an important side note that should be made is that not everyone liked hearing
the heartbeat of the person that was talking to them. As said earlier, this may only have been
the case when the heart rate contradicted the facial expressions, but since we did not check
for this we can only speculate. Another explanation is that people are not used to this way of
communication. Perhaps, after a little more practice, the heartbeat sound would appear more
natural, which would imply that it is simply something people need to get used to first.
Chapter 4

General Discussion

4.1 Summary

In this thesis we have presented our work on the perception of affective signals by both computers and humans. In the first part of our research we investigated the effectiveness of different modalities — video (for facial expressions), audio (for vocal expressions) and physiology — and every combination of these modalities regarding their usefulness for automated emotion recognition. Five emotion classes were used — happy, relaxed, sad, angry, neutral — which were induced using the method of autobiographical recollection (Appendix A). The extraction of features was done for each modality separately, after which they could be combined. For combining the feature vectors a fusion method named feature-level fusion was used, meaning that the feature vectors from the to-be-combined modalities were concatenated, resulting in joint multi-modal feature vectors. In order to find the most useful features, and to remove the most noisy ones, feature selection was applied. Then, classification was performed, using two classifiers, a Support-Vector Machine (SVM) and a Multi-Layer Perceptron (MLP). Three classification methods were applied, Leave-one-Out (LVO): training on the data of all but one subjects and testing on the left-out subject; Cross Validation per Subject (CVS): 5-fold cross validation applied to the data of all subjects separately; and Cross Validation Overall (CVO): 5-fold cross validation applied to the combined data of all subjects. For the MLP classifier, only the latter method was applied. The results revealed that (1) physiology appeared to be a very useful modality, (2) the most successful modality combination turned out to be video + physiology, and, perhaps most importantly, (3) combining information from multiple modalities almost always led to an increase in classification performance when compared to using a single modality on its own.

In the second part of our research we investigated the effect of an amplified physiological signal, namely a heartbeat, on people’s judgment of a person’s experienced emotional intensity. We conducted an experiment in which participants looked at videos of persons telling about an emotional event from their life. The event was either a neutral one, or one where participants experienced anger. The videos were accompanied by a heartbeat sound that was either slow or fast. Participants were told that the heartbeat sound was that of the person in the video, but in fact it was an artificial heartbeat. The audio of the telling person was muted, so the participants’ judgments would only be based on facial expressions and heart rate. The videos were presented in pairs. For judging emotional intensity participant used a double Visual Analogue Scale (VAS) on which they could indicate their emotional intensity judgment for the persons in both videos. The results revealed a significant effect of both facial expressions and heart
rate on the judgment of emotional intensity. When facial expressions and heart rate were in agreement, they appeared to strengthen one another effect. In case of disagreement, subjects appeared to have taken a rough average in order to concede to the information coming from both modalities. From the questionnaire, asking the participants how they felt about hearing the heartbeat sound, there did not seem to be much consensus about whether they liked it or not.

In sum, it appears that for both computer and human, the judgment of emotions can be greatly influenced (improved) by physiological information. Moreover, for both, facial and physiological information seem to complement each other quite successfully. In Chapter 1 we mentioned two problems that Sebe et al. (2004) pointed out about (the lack of) affect processing by computers. The first problem was that, when replacing a human in a human-human interaction by a computer, the communication of affect would disappear, with the unfortunate result that the human’s potential frustration, anger, boredom, etc. would be completely ignored. The second problem was about the great amount of affective information that gets lost when two people communicate via computers (e.g. e-mail or instant messaging), instead of face-to-face. With the work presented in this thesis we could be one step closer to solving both of these problems.

4.2 Applications

As mentioned in Chapter 1, a computer that could recognize and appropriately respond to a user’s affective states could greatly improve the HCI experience for users. Applications that come to mind are for example ambient environment that can change light conditions for calming down a stressed person, computers that can deal with a user’s frustration and give the user emotional support. Picard and Klein (2002) argue that no deep knowledge of human emotions is needed in order to be able to give emotional support (e.g. a dog is capable of doing this). Bickmore and Picard (2004) showed that a computer can give a sense of feeling cared for, using text based conversation combined with an avatar. A computer’s ability to recognize a user’s affective state could also improve remote communication between humans by transferring information about the other person’s affective state, which would be naturally perceived in a face-to-face interaction.

4.3 Future work

The experiment described in this chapter enabled us to answer some key questions about the influence of a physiological signal in the field of human-human interaction. But the results led to several new questions, which could be interesting for future research on this topic. From the questionnaire that was given to participants after the presentation, asking them about how they felt about hearing the heartbeats of the persons in the videos, it appeared that several participants experienced the heartbeat sound as disturbing. This led to the question: Do people who find the heartbeat sound disturbing, always find it disturbing, or only in case it is in conflict with the facial expressions (e.g. fast heartbeat in combination with neutral facial expressions)? Before thinking about any future applications, it would be good to have this question answered first. Another interesting point is the fact that, in the experiment, audio of the person in the video was muted, in order to prevent participants from making their judgment on the story that was being told. However, according Mehrabian (1968), only about 7% of the effect of a message is caused by verbal information (i.e. what is said), whereas the other 93% is caused by facial and vocal information. Therefore, it may be interesting to study the effect of facial and physiological signals on emotion(al intensity) judgment, when these are combined with verbal information in order to see if these numbers can be validated.
4.4 Conclusion

The field of affective computing is still in its infancy. It is however, a field that is gaining more and more attention since the value of a computer’s ability to recognize and express emotions is becoming better understood. There is still much work to be done in this area. From the work presented in this thesis it became clear that interaction with and via computers does not necessarily mean a restriction of affective signals. Computers can even be used to magnify certain signals — such as physiological signals — that would be harder to be perceive in direct face-to-face communication. For automated emotion recognition we have shown that combining information from multiple modalities can greatly improve the recognition performance. Moreover, the use of physiological information has proven itself to be very effective for both automated emotion recognition as well as for emotion judgment by humans. These finding enable us to conclude that taking into account physiological information can enrich both the recognition and the communication of humans’ affective states.
References


Appendix A

Data gathering

To be able to test how well humans and machines are able to recognize emotions, data is necessary to test them on. Hence, the first step of this research was a trial in which we collected emotional self-reports, physiological recordings, speech, and facial expressions from different people in different emotional states. As discussed before, it is key to generate ecologically valid emotional responses so that our results can be generalized to situations outside the lab. For this, we employed autobiographical recollection because it is a very strong emotion inducer (Levenson, Carstensen, Friesen, & Ekman, 1991; Zaki, Bolger, & K. Ochsner, 2009) and is likely to generate more ecologically valid emotional responses compared to acted emotions that are sometimes used for emotion recognition research (e.g., Picard et al., 2001). Furthermore, it involves speech whereas many passive induction techniques do not involve the participants having to speak.

A.1 Participants and Design

Participants were eight women and nine men aged 18 to 26 (M = 21.1) who received 25 euro for their participation. All participants were undergraduates at a Dutch university and were native Dutch. None of the participants reported any cardiovascular problems. The within-participant factor Emotion type (Happy, Sad, Angry, Relaxed, and Neutral) was counterbalanced over the participants.

A.2 Materials

Physiological recordings were done with a Mobi-8 of TMS International b.v. Electrodermal activity (EDA) was measured with two Velcro strips with dry electrodes strapped around the distal phalanx of the index and ring finger of the non-dominant hand. The EDA signal was sampled with 128 Hz. Skin temperature (ST) was measured with a thermistor strapped with mediacal adhesive tape to the distal phalanx of the little finger of the non-dominant hand. The ST signal was sampled with 128Hz. Respiration was measured with a gauge band positioned over the clothes around the chest at a sampling frequency of 128Hz. An electrocardiogram was taken at 1024Hz using Ag/AgCl electrodes on the standard Lead-II placement. Facial expressions were recorded using a Logitec Quickcam webcam at 25fps with a resolution of 640 x 480 pixels. The webcam was positioned at the top of the computer screen that the participants were facing. Audio was recorded using a Sennheiser MZT 100 microphone positioned on the desk the participants were sitting at.
The entire experiment was done on a computer with the participants sitting behind a desk that held the computer screen, keyboard and mouse, microphone and webcam. The experimenter was in a control room next to the room the participants were in.

A.3 Procedure

After arriving in the lab, participants signed an informed consent form. In this informed consent, we explicitly told them that their data would not be shown to other people to make sure this would not limit their emotional expressions. Only at the end of the experiment did we debrief them about the true intentions we had for using this data and there we asked them permission to show the recordings we made to other people. One participant did not consent to this and was therefore excluded from the human emotion recognition experiment described later on.

In the first phase of the experiment, participants were asked to recall two events from their personal experience that made them feel extremely happy, relaxed, sad, and angry. They were also asked to recall getting up this morning and going home last evening to provide two possibly neutral events. Subsequently, they were asked to write one paragraph about two events for each of the five emotion types, shortly describing the event and their feelings during the event. Participants provided a title and rated each event on emotional intensity (very weak to very intense), valence (very unpleasant to very pleasant), and arousal (very relaxed to very aroused) on nine point Likert type scales. This took 30 to 45 minutes. In the second phase, the experimenter attached the physiological sensors and checked by the signals visually. Subsequently, two potentially interesting individual difference measures were administered. First, the participant filled out the Berkeley Emotional Expressivity Questionnaire (Cronbachs a = .88), a 16-item questionnaire assessing emotional expressivity (Gross & O. P. John, 1997). Second, we assessed participants cardiac awareness (which is an objective index of emotional expressivity; Herbert, Pollatos, Flor, Enck, & Schandry, 2010) by having them count their number of heart beats without moving and with their hands beside each other on the desk. Participants counted their heart beats over 25, 35, and 45 seconds without knowing the duration of these periods. Cardiac awareness was defined as

\[
\frac{1}{3} \sum \left(1 - \left(\frac{|recordedHBs - countedHBs|}{recordedHBs}\right)\right)
\]

Where the sum was taken over the three periods of counting. In the meantime, the experimenter selected one event for each of the five emotion types. The selection was based on the ratings of the participants and their description of the event.

In the third phase, we used autobiographical recollection to induce the five emotions in the participants. For each of the five emotions, the participants first watched a five minute aquatic movie to make sure they started each condition in the same neutral state (Piferi, Kline, Younger, & Lawler, 2000). After the baseline video, the participant turned over a piece of paper with the title of one of the events they described earlier. They were instructed to take a minute to recall the events and try to relive their feelings during the event. When they felt they were ready after about one minute, they pressed continue and described the event and their feelings during the event in two to three minutes. After each description they were told to rate their feelings during the event disclosure on emotional intensity (very weak to very intense), valence (very unpleasant to very pleasant), and arousal (very relaxed to very aroused) on nine point Likert type scales.
Instructions explicitly stated that they had to rate their feelings during the disclosure of the event and not during the event itself.

In the fourth and final phase, after all five events were described, participants received (both audio and video of) each of the recordings we made of their disclosures. While watching the video, participants had to continuously rate how they felt during the recording on a 9 point Likert type scale using the left and right arrow key. This is a validated method of emotional state assessment (Levenson & Ruef, 1992). Half of the participants first rated their arousal on all five movies and then their valence, whereas this was done vice versa for the other half of the participants.

References


Appendix B

List of extracted facial features

Table B.1: List of extracted facial features, using the OKAO vision software package.

<table>
<thead>
<tr>
<th>Confidence</th>
<th>Left Outer Eye Corner X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Face Average X</td>
<td>Left Outer Eye Corner Y</td>
</tr>
<tr>
<td>Face Average Y</td>
<td>Left Inner Eye Corner X</td>
</tr>
<tr>
<td>Face Standard Deviation</td>
<td>Left Inner Eye Corner Y</td>
</tr>
<tr>
<td>Right Eye Ratio</td>
<td>Right Inner Eye Corner X</td>
</tr>
<tr>
<td>Left Eye Ratio</td>
<td>Right Inner Eye Corner Y</td>
</tr>
<tr>
<td>Mouth Ratio</td>
<td>Right Outer Eye Corner X</td>
</tr>
<tr>
<td>Face Pitch</td>
<td>Right Outer Eye Corner Y</td>
</tr>
<tr>
<td>Face Yaw</td>
<td>Right Upper Lip X</td>
</tr>
<tr>
<td>Face Roll</td>
<td>Right Upper Lip Y</td>
</tr>
<tr>
<td>Right Pupil X</td>
<td>Left Upper Lip X</td>
</tr>
<tr>
<td>Right Pupil Y</td>
<td>Left Upper Lip Y</td>
</tr>
<tr>
<td>Left Pupil X</td>
<td>Right Lower Lip X</td>
</tr>
<tr>
<td>Left Pupil Y</td>
<td>Right Lower Lip Y</td>
</tr>
<tr>
<td>Upper Lip Center X</td>
<td>Left Lower Lip X</td>
</tr>
<tr>
<td>Upper Lip Center Y</td>
<td>Left Lower Lip Y</td>
</tr>
<tr>
<td>Lower Lip Center X</td>
<td>Gaze Tilt</td>
</tr>
<tr>
<td>Lower Lip Center Y</td>
<td>Gaze Pan</td>
</tr>
<tr>
<td>Right Mouth Corner X</td>
<td>Left Eye Openness Level</td>
</tr>
<tr>
<td>Right Mouth Corner Y</td>
<td>Right Eye Openness Level</td>
</tr>
<tr>
<td>Left Mouth Corner X</td>
<td>Mouth Openness Level</td>
</tr>
<tr>
<td>Left Mouth Corner Y</td>
<td></td>
</tr>
</tbody>
</table>

46
Appendix C

Instruction form of the Emotional Intensity Judgment experiment

Informatie onderzoek Emotion Intensity Judgment

In dit document staat informatie over de procedure van het experiment. Lees de volgende paragrafen aandachtig door.

1. Doel van het onderzoek

Het doel van dit onderzoek is om te kijken hoe mensen de intensiteit van emoties beoordelen.

2. Experiment

Tijdens het experiment krijgt u video’s te zien van personen die vertellen over een emotionele gebeurtenis in hun leven. Het geluid is uitgeschakeld, zodat u niet kunt horen wat de persoon in de video zegt. Tijdens de video wordt de hartslag van de vertellende persoon afgespeeld. Het experiment bestaat uit 10 delen waarin telkens twee video’s worden afgespeeld. Ieder video-paar wordt drie keer herhaald. Nadat u een video-paar drie keer gezien heeft wordt u gevraagd de intensiteit van de emotie van de personen in beide video’s te beoordelen, door middel van een schaal die loopt van 1 tot 9. De onderzoeker zal gedurende het gehele experiment bij u zijn en beschikbaar zijn voor vragen en opmerkingen.

3. Risicos en neveneffecten

Het experiment is gebaseerd op aanwezige kennis van de onderzoeker en is veilig en pijnloos voor de proefpersonen. Door mee te doen aan dit experiment zal u geen risicos ondervinden die anders zijn dan de risico’s die u normaal tijdens het dagelijks leven ondervindt.

4. Beindiging van het experiment
U heeft het recht om op elk moment het experiment te beindigen zonder opgaf van reden. De deelname is volledig vrijwillig en zonder verplichtingen. Er zijn geen nadelen voor u verbonden aan beindiging van het experiment uwerzijds.
Appendix D

Adapted Multi-Dimensional Emotional Empathy Scale
Table D.1: Adapted Multi-Dimensional Emotional Empathy Scale (Caruso & Mayer, 1998) for measuring empathy. The scale was modified from a 5-point to a 9-point scale.

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
<th>Strongly Disagree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I feel like crying when watching a sad movie.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Certain pieces of music can really move me.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>Seeing a hurt animal by the side of the road is very upsetting.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>I don’t give others’ feelings much thought.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>It makes me happy when I see people being nice to each other.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>The suffering of others deeply disturbs me.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>I always try to tune in to the feelings of those around me.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>I get very upset when I see a young child who is being treated meanly.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>9</td>
<td>Too much is made of the suffering of pets or animals.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>If someone is upset I get upset, too.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>When I’m with other people who are laughing I join in.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>It makes me mad to see someone treated unjustly.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>I rarely take notice when people treat each other warmly.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>I feel happy when I see people laughing and enjoying themselves.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>15</td>
<td>It’s easy for me to get carried away by other people’s emotions.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>16</td>
<td>My feelings are my own and don’t reflect how others feel.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>17</td>
<td>If a crowd gets excited about something so do I.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>I feel good when I help someone out or do something nice for someone.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>19</td>
<td>I feel deeply for others.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>20</td>
<td>I don’t cry easily.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>21</td>
<td>I feel other people’s pain.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>22</td>
<td>Seeing other people smile makes me smile.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>23</td>
<td>Being around happy people makes me feel happy, too.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>24</td>
<td>TV or news stories about injured or sick children greatly upset me.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>I cry at sad parts of the books I read.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>26</td>
<td>Being around people who are depressed brings my mood down.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>27</td>
<td>I find it annoying when people cry in public.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>28</td>
<td>It hurts to see another person in pain.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>29</td>
<td>I get a warm feeling for someone if I see them helping another person.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
<tr>
<td>30</td>
<td>I feel other people’s joy.</td>
<td>1 2 3 4 5 6 7 8 9</td>
<td>9</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Pipeline of the automated emotion recognition process.</td>
<td>8</td>
</tr>
<tr>
<td>2.2</td>
<td>Valence/arousal model with 5 emotions. Based on (Russell, 1980).</td>
<td>8</td>
</tr>
<tr>
<td>2.3</td>
<td>Creation of binary code for each pixel. Adopted from (Ahonen et al., 2004).</td>
<td>10</td>
</tr>
<tr>
<td>2.4</td>
<td>Concatenation of LBP histograms from every region of the face area. Adopted from (Shan et al., 2009).</td>
<td>10</td>
</tr>
<tr>
<td>2.5</td>
<td>Example of feature point tracking by OKAO. Adopted from (Ahn et al., 2009).</td>
<td>11</td>
</tr>
<tr>
<td>2.6</td>
<td>The five emotional states positioned on the valence/arousal space according to the continuous valence- and arousal ratings given by the subjects.</td>
<td>13</td>
</tr>
<tr>
<td>2.7</td>
<td>Classification results of CVO method using the SVM classifier. The results are sorted in ascending order. $p + v + a$ stands for physiology + video + audio.</td>
<td>18</td>
</tr>
<tr>
<td>2.8</td>
<td>Classification results of CVO method with the MLP classifier. The results are sorted in ascending order. $p + v + a$ stands for physiology + video + audio.</td>
<td>21</td>
</tr>
<tr>
<td>3.1</td>
<td>Double visual analogue scale (VAS) for judging emotional intensity of Person A and B.</td>
<td>30</td>
</tr>
<tr>
<td>3.2</td>
<td>Experimental procedure.</td>
<td>31</td>
</tr>
<tr>
<td>3.3</td>
<td>Emotional intensity ratings per facial expression + heart rate combination.</td>
<td>32</td>
</tr>
<tr>
<td>3.4</td>
<td>Scatter plot showing the average difference between the emotional intensity scores for angry and neutral facial expressions for every subject, plotted against their emotional empathy score.</td>
<td>34</td>
</tr>
<tr>
<td>3.5</td>
<td>Scatter plot showing the average difference between the emotional intensity scores for fast and slow heart rates for every subject, plotted against their emotional empathy score.</td>
<td>34</td>
</tr>
</tbody>
</table>
### List of Tables

2.1 Brief description of the extracted physiology features .......................... 12
2.2 An overview of the classification results for the SVM classifier, for every modality combination and three classification methods, Leave-One-Out (LVO), Cross Validation per Subject (CVS) and Cross Validation Overall (CVO). .................. 17
2.3 The resulting features using best-first search for subset selection. .......... 18
2.4 Confusion matrix of 5-fold cross-validation result with the SVM classifier, using best-first search feature subset. ................................. 19
2.5 Classification results of 5-fold cross-validation with the SVM classifier for different physiology sub-modalities. ................................. 19
2.6 Confusion matrix of SVM classification result using re-computed emotion classes based on the overall valence- and arousal ratings. ................. 20
2.7 An overview of the classification results for the MLP classifier for every modality combination ......................................................... 20
2.8 Confusion matrix of of 5-fold cross-validation results with the MLP classifier, using the physiology + video combination. ......................... 21
2.9 Confusion matrix of MLP classification result using re-computed emotion classes based on the overall valence- and arousal ratings. ................. 22
3.1 The results of the emotional intensity judgment experiment, showing the mean, standard deviation, minimum and maximum of the average rating values per subject for each facial expression + heart rate combination. ................. 33
B.1 List of extracted facial features, using the OKAO vision software package. ...... 46
D.1 Adapted Multi-Dimensional Emotional Empathy Scale (Caruso & Mayer, 1998) for measuring empathy. The scale was modified from a 5-point to a 9-point scale. 50