Creating a Mood Database for automated affect analysis

Computer Science

Interactive Intelligence Group

Faculty of Electrical Engineering, Mathematics and Computer Science

Joris Albeda

August 2016

TUDelft

Creating a Mood Database for automated affect analysis

By

Joris Albeda

1514172

in partial fulfilment of the requirements for the degree of

Master of Science

in Computer Science

at the Delft University of Technology,

to be defended publicly on September 27th, 2016

Supervisor:	Dr. J. A. Redi,	TU Delft
Thesis committee:	Prof. Dr. C. M. Jonker,	TU Delft
	Dr. J. A. Redi,	TU Delft
	Dr. Ir. A. Bozzon,	TU Delft

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Preface

As a student of both the TU Delft and the Acting School of Rotterdam, I envisioned this final project of the TU Delft as a way to bring both interests together: using the art of acting together with the science of computing to bring a machine closer to understanding the human concept of a mood. This work is the end result of that endeavour.

I would like to thank my girlfriend, Linda van Lierop, for her endless love, comfort and support, for her suggestions and her illustration on the front cover to this report. Likewise, I wish to express my gratitude to my parents, Monique and Peter Albeda, and my sisters, Laurien and Inger, for their constant support throughout this project.

I would like to show my gratitude to Judith Redi for her feedback and guidance during the writing of the thesis, and for supervising the project. Furthermore, I am very grateful to Christina Katsimerou for providing the daily supervision for the project and her help in realising this work.

Additionally, my thanks go out to Tina Wrede, for helping me overcome my doubts and see this project through to the end.

My special thanks go to the fifteen actors on whose talents I relied to create the raw data for this database:

Kelly de Korte
Ylva Rietman
Thijs Ringelberg
Nadine Febranof
Cheyenne de Vink
Daphne Heij
Ivo van der Hoeven
Daniel van Dijk
Chloé Melgarejo
Ashley Boom
Tim Bruyn
Tom Strik
Lars Jordens
Melissa Breedveld
Inger Albeda

Finally, I wish to extend my heartfelt thanks to my friends, some from Rotterdam, some from Delft and many from elsewhere, for lending me support and encouragement when needed.

And thank you, reader, for reading this thesis. Without you, these words would be nothing more than words.

Joris Albeda
August 2016

Abstract

Affect-adaptive systems are dependent on their ability to automatically recognize a user's affective state. This study aims to contribute to the creation of an affect-adaptive system that can recognize negative moods of elderly in care homes from a video feed, and improve it by adapting the lighting in the room. An affective database of videos portraying different moods is required to train such a system. While many affective databases exist already, they are primarily targeting emotions rather than mood. Therefore, we introduce a new database of annotated videos that can be used for mood recognition. To maintain control over which moods are depicted in the videos in the database, we combine the use of mood induction and acted performance to portray the moods in a realistic way, incorporating in the acted scripts the results from a series of interviews with caretakers in care homes. The database covers three visual modalities: body, face and 3D Kinect data for a total of 24 hours of recorded video material. We use crowdsourcing to annotate such a large amount of material in terms of perceived mood of the person portrayed in the videos, by outsourcing via the internet the annotation task to a large number of paid annotators. A risk of using crowdsourcing is unreliable annotator performance, due to the low level of control applicable to the annotation process. We deal with this problem by filtering the annotations according to predefined criteria, checking for task commitment and self-consistency of the annotators. We validate our use of the combination of induction and actors with a comparison between the intended mood, the mood felt by the actors, and the mood perceived by annotators. Furthermore, we demonstrate that crowdsourcing is a promising tool for the annotation of mood.

Contents

Chapter 1: Introduction

A natural part of being human is experiencing affective states. We express our *emotions*, our short-term affective states (Russell and Feldman-Barrett 1999), to make ourselves understood. The experience of an emotion produces a *feeling* (Ketai 1975). Our *moods*, our long-term affective states, impact how we think (Russell and Mehrabian 1977). *Affect* includes each of these phenomena. Specifically, affect is "*an encompassing term which includes emotions, feelings, motivational impulses, and moods together*" (Berking and Whitley 2014).

Developments in human-computer interaction (HCI) over the last few decades have focused on user-centred technology design, increasing the importance of making the interaction between a computer and its user as natural as possible. (Jacko 2012). Studies have shown that many social and natural factors in interaction between humans are relevant for interaction between human and machine as well (Reeves and Nass 1996). For the purpose of creating more natural interactions between computers and users, research has sparked to integrate affect in HCI (E. 2003).

A specific appliance of affect in HCI is the automatic recognition of affective states. The ability to recognize and express affective states is vital for successful communication (G, et al. 2010). Affect recognition is an inherent skill for humans from which computer systems can greatly benefit. A system can learn from the feedback retrieved from an affective expression, or change its behaviour based on the affective states it detects (R. W. Picard 2000). Examples of applications of affect recognition include tutoring systems (Forbes-Riley and Litman 2009) (Liping Shen 2009), pervasive systems in homes (Varshney 2007) (Ramos 2008) and user interfaces (Duric, et al. 2002). A specific purpose of automatic affect recognition is to *adapt* the system to the user's affective state. Systems with such a purpose are known as *affect-adaptive systems*. After the system has sensed and recognized the user's affective state, the system can decide how to react. The objective could be to change the user's affective state, to maintain the affective state to optimize the user's performance in a task or to adapt to the affective state in order to increase how enjoyable the user experience is (Hudlicka 2003).

Automatic affect recognition has become an active research area within computer science over the last few decades (Z., et al. 2009) (B. and J. 2003). There are several ways for a system to automatically recognize an affective state, e.g. by analysing physiological measurements, visual data (facial and/or bodily expressions) or audio feeds (typically speech) (Hudlicka 2003). Automated emotion recognition based on the analysis of such signals has achieved success rates matching those of humans: 80% using psychophysical measurements (Picard, Vyzas and Healey 2001), 80% using facial expressions (I., et al. 2003) and 60% using speech (Hudlicka 2003). However, while there is plenty of previous work focussed on *emotion* recognition, there is much less existing research on automatic *mood* recognition.

This thesis contributes to the creation of an affect-adaptive system that aims to improve a negative mood upon recognizing it. Such a system requires a mood recognition system, which typically relies on supervised machine learning (D'mello and Kory 2015). This learning requires example input that can be used to establish ground truth. Thus, this system needs a collection of annotated visual mood data. Such a collection is known as a mood database

(Z., et al. 2009). This work presents the creation of a mood database. This introduction explains its scope, its original contribution and the structure of the document.

1.1 Problem statement

1.1.1 Application context

This work is a contribution to the ACE project (Adaptive Ambience Creation in Care Centres for Elderly), which aimed at creating intelligent systems that can improve an elder person's mood by dynamically changing the surrounding lighting (Kuijsters, et al. 2015). Residents in a care centre often need time to adjust to their new home and lifestyle, which can cause anxiety (Nay 1995). They often miss their relatives and old life, potentially leading to gloominess and sad moods (Lee, Woo and Mackenzie 2002). Especially new residents often experience such negative moods. The ACE project aims to increase the welfare of these residents with the possibilities that affective computing offers. The objective is to create an affect-adaptive system that can react to these negative moods and influence them positively by adjusting the lighting. This system will have to recognize mood in an unobtrusive way, as it will likely deployed in non-interactive daily-life settings (e.g. where the resident is alone in his/her room), and it would be inconvenient to the residents if the recognition system would interfere in their daily lives, e.g. by requiring them to keep sensors on their bodies.

1.1.2 A mood recognition system

Current work on affect recognition is largely based on *emotion* recognition. However, for this system, it is important to react to the long-term affective state of the user, the *mood*. As emotions are shorter and more intense than mood (Russell and Feldman-Barrett 1999), a system reacting to emotion would make sudden and frequent changes to the lighting, which could be very unpleasant to the user. Therefore, we focus our research on mood recognition. As much of the literature covers affect recognition in general rather than emotion or mood recognition in particular (Tao and Tan 2005) (Z., et al. 2009), this study draws from insights from affect recognition in general as well as specifically mood recognition.

Affective states are psychological phenomena that cannot be directly observed but rather interpreted from communication and context and influenced by cultural influences (Elfenbein and Ambady 2002), making affect recognition a complex problem. As the process of automatically detecting and recognizing an affective state requires the development of a classifier or a regressor, affect recognition is a pattern recognition task (D'mello and Kory 2015). Generally, affect recognition consists of three parts: processing the input, extracting information on the expression (e.g. facial expression, bodily movement or gestures), and classifying the expression, with either supervised or unsupervised learning (Pantic and Rothkrantz 2000) (Kleinsmith and Bianchi-Berthouze 2013) (De la Torre and Cohn 2011). Supervised learning is the most common method (De la Torre and Cohn 2011). Many techniques have been used for the classification (Tao and Tan 2005), such as neural networks (Kulkarni, Reddy and Hariharan 2009), Hidden Markov Models (HMM) (Mitra and Acharya 2007), and support vector machines (SVM) (Bartlett, et al. 2004). With supervised learning, these techniques rely on a classifier that needs to be trained using example input. In the case of a system that can perform mood recognition based on visual input, the classifier is trained using a collection of annotated visual mood data. The machine learning is the task of creating a mapping from the input data, the visual material, to the expected output, the annotations. This collection of visual data and annotations is known as a mood database (Z., et al. 2009).

Many affective databases have been proposed to the purpose of training and benchmarking automatic affect recognition systems (Douglas-Cowie, et al. 2007) (Busso, et al. 2008) (A. e.

Metallinou 2010), of which we will give an overview in Chapter 2. Most of these databases are centred on emotion. This study, however, is focussed on a system that is able to recognize and respond to long, subtle affective states in a non-interactive setting. Due to the inherent differences between emotion and mood (Beedie, Terry and Lane 2005), emotional data would be ill suited as training data for mood recognition. We will cover these differences in Chapter 2. With the lack of usable training data for mood recognition, a new mood database is necessary.

1.1.3 An annotated mood database

An affective database consists of two core components (Cowie, Douglas-Cowie and Cox 2005): the *material* that portrays the affect (in the form of videos, images or signals from any kinds of sensors), and the *affective annotation* of this material, in the form of quantitative or qualitative labels that describe its affective content.

For the creation of the material, three key issues must be considered (Douglas-Cowie, et al. 2007). The first of these is the *scope*, which involves the type of material that is recorded, the range of the affective states that are shown, the diversity in the people expressing these affective states and the modalities through which they are expressed (such as body, face and gestures). These modalities, or expressive channels, are related to how the affective states are captured; while most databases use visual data on the body and face captured through cameras and/or speech recorded with a microphone, some databases include more advanced sensors such the CALLAS database, which includes sensory data from a Wii controller (Caridakis, et al. 2010). The second issue is the *naturalness* of the affective states. These affective states may be acquired from a source where the states occur naturally, they may be induced in a laboratory environment, or they may be portrayed by actors. While the latter two methods provide more control over the resulting content, it is a non-trivial task to acquire realistic portrayals using these methods (Cowie, Douglas-Cowie and Cox 2005). As will be discussed in Section 2.2.1 , the issue is a trade-off between naturalness and control (Sneddon, et al. 2012). The design choices will impact how realistic the displayed affect will be, and to what extent the researcher can be certain that data displays the affect intended by the researcher. Finally, a choice must be made concerning the *context* of the material. The affective states may occur in an interactive setting, they may or may not have a clear cause, and they may be affected by the scenario in which they take place. When considering these design decisions, we can gain insights from existing work; however, since the existing work largely concerns emotions rather than mood, we cannot base our work entirely on it.

The annotation of the material is a fundamental part of an affective database (Douglas-Cowie, et al. 2007). For the database to be used as training data, it must be clear which moods the material is intended to train the system for. However, due to the subjective nature of mood, it is impossible to fully control the moods that are portrayed in the material. Hence, to be able to pair each item in the database with the corresponding mood portrayed in it, researchers often resort to ask external observers to report the mood they perceive in the content, as second best approximation, given the well-developed empathic skills of humans (Douglas-Cowie, et al. 2003). These skills are, however, not equally developed in each person. Hence, the perception of mood is a subjective matter (Siegert, Böck and Wendemuth 2014). Combining the opinions of multiple observers helps to overcome the problem of the subjectivity of the perception of a single viewer (Vidrascu and Devillers 2005). Such techniques, which will be further discussed in Section 2.2.2 are used with the intention of generating *reliable* annotations. Reliability “concerns the extent to which measurement is repeatable and consistent – that is, free from random error” (Calvo, et al. 2014). It is crucial that the annotation process produces reliable annotations that can be used to train an affect recognition system effectively.

Annotation of an affective database is traditionally done by a small number of experts (Busso, et al. 2008). However, this is often an expensive and time-consuming task, especially considering the large size of the intended database. We look towards *crowdsourcing* as an alternative, a means for solving problems online. The approach for crowdsourcing is to divide a large task into smaller tasks and outsource these smaller tasks on the internet (Howe 2006). Its functionality will be further explained in Section 2.3.1 .

As will be discussed in Section 2.3.3 , reliability is an issue when using crowdsourcing for affective annotation (Nowak 2010). Some of the annotators recruited through crowdsourcing (also known as workers) are not motivated or are unable (due to, e.g., insufficient understanding of the instructions) to carry out the task properly, negatively influencing the quality of their work (Eickhoff and Vries 2013). Furthermore, since these workers are typically not trained to perform affective annotation, using crowdsourcing for this application may result in noisy annotations (Hsueh, Melville and Sindhwani 2009). Because of this, it is vital that the design of a crowdsourcing task incorporates mechanisms to acquire only high quality results. As mentioned earlier in this section, the reliability of affective annotations is the degree to which the annotating process is a consistent measurement, with as little random noise in the annotations as possible (Calvo, et al. 2014). Because of this, the reliability of crowdsourced annotations is typically measured as the agreement between the annotators (Nowak 2010).

This thesis will address the abovementioned major objectives: acquiring material that realistically portrays the intended mood, and obtaining reliable affective annotation for that material.

1.2 Research questions

The main objective of this project is to create an annotated mood database that can be used for an affect adaptive system that can accurately recognize the mood of elderly. To achieve this goal, four research questions have been formulated for the project, where the first two are related to the acquisition of representative material, and the last two involve the reliability of crowdsourced annotations.

RQ1. What are the defining features of the moods depression and anxiety when experienced by elderly?
RQ2. How can videos be obtained that accurately portray depression and anxiety?
RQ3. How can the reliability of crowdsourced annotations be controlled effectively?
RQ4. Is crowdsourcing an appropriate tool to obtain accurate annotations for a mood database?

1.3 Contribution

The contribution of this project can be described in three parts. Firstly, we created a new affective database containing visual data showing the moods we want the system to recognize in the setting of a care home. To incorporate the context of a care home and to control the displayed mood, we used actors to portray the desired moods. In order to increase the naturalness of the database, we performed mood induction (Westermann, Stahl and Hesse 1996) on these actors. The product of this project is a new database specifically created to display mood states (especially experienced by elderly), which can be used as suitable material for the research of mood recognition.

Secondly, we built an annotation tool that could be used to annotate the material via crowdsourcing. This tool is based on the music annotation tool made by Soleymani (Soleymani, Caro, et al. 2013). We adapted his tool to annotate videos instead of music, in terms of mood in addition to emotion. The source code will be provided on request.

Implemented in HTML and JavaScript, this tool can be used in the future by other researchers to use crowdsourcing for the annotation of videos.

Finally, we explored the limitations of crowdsourcing for collecting affective (mood) annotations of videos in terms of reliability. Based on the literature (Eickhoff and Vries 2013) (Hoßfeld, Hirth, et al. 2014), we devised a number of mechanisms to control the reliability of the resulting annotations and implemented them into our tool. These mechanisms will be explained in Chapter 5. We analysed the effects of these mechanisms. Based on the analysis, we provide means to keep the reliability of crowdsourced annotations as high as possible in future work.

1.4 Thesis outline

The remainder of this thesis is organized as follows. In Chapter 2, we explore the relevant literature concerning mood and emotion, mood databases and crowdsourcing. We then describe the methodology in Chapter 3. We go into detail on the creation of the database in Chapter 4. We then describe how the database was annotated using crowdsourcing techniques in Chapter 5. Chapter 6 presents our analysis on the acquired annotations, and on the reliability mechanisms. Finally, we draw conclusions and provide recommendations for future work in Chapter 7.

Chapter 2: Literature review

This chapter presents the literature study that forms the basis for this project. We first investigate the field of affect recognition, and specifically of the type of data that is used to train automatic affect recognition systems, i.e., the affective databases. There are many examples of affective databases (Douglas-Cowie, et al. 2007) (Busso, et al. 2008) (A. e. Metallinou 2010), the vast majority of which focus on portraying *emotions*. The purpose of this work is to create a *mood* database, which is a different affective state from emotion. Hence, the work in this thesis can be inspired by existing affective databases, but also needs to substantially depart from them.

In this review, we first discuss the difference between mood and emotion and the implications this has on the decisions in our project. We then introduce the choices and issues that are integral to the design of an affective database. Finally, we discuss crowdsourcing as a tool for the annotation of affective data. We describe the basic functioning of crowdsourcing, motivate the use of crowdsourcing specifically for annotation, provide examples of prior work using crowdsourcing techniques for this purpose and discuss the issues concerning reliability when using crowdsourcing.

2.1 Affect and its representation

2.1.1 Mood and emotion: different affective states

Russell (Russell and Mehrabian 1977) describes mood as being a *prolonged core affect with no object or with a quasi-object*, commenting on its fuzziness and lack of definition in terms of duration or stability. In short, an emotion is an affective reaction to a certain stimulus, whereas mood is an affective state with longer duration. Furthermore, emotion affects behaviour and physiology, whereas mood affects an enduring state of affect and cognition (Rottenberg 2005). The time between a mood and its cause tends to be greater than the time between an emotion and its elicitation (W. Morris 1992).

Beedie et al (Beedie, Terry and Lane 2005) investigated differences between emotion and mood by comparing views on the subject among a non-academic population with views in the literature. They found a number of relevant results. They agree with Russell that mood is connected to thought processes and cognition, whereas emotion has physical causes; also, mood endures longer than an emotion. Additionally, mood is more personal than emotion, and easier to hide. Finally, it is considered more difficult for someone to describe their current mood than their current emotion.

Ekman (P. Ekman 1999) claims that moods have no own distinctive signals. Instead, we infer mood from the signals of the emotions that we associate with the mood, at least in part. For example, we might deduce that someone is in a cheerful mood because we observe behaviour that matches a joyful emotion, which we associate with a cheerful mood. Thus, while mood and emotion are indeed distinct, it is worthwhile to investigate connections between the two in practice.

2.1.2 Affect representation and measurement

To be able to collect, process and use affective information in HCI and especially in affect adaptive system, it is important to be able to describe affective states in a representative way. In this section, we will cover methods of representing affect, as well as tools that are commonly used to measure affect.

2.1.2.1 Affect Representation Models

Two common models for affect representation are the categorical and dimensional models. These two are most widely used as reference for computational models for the task of affect recognition (Z., et al. 2009). Both emotion and mood can be represented with these models (Hu 2010) (Fessl, et al. 2012).

The categorical system describes an affective state by assigning it to one of a set of categories. In the case of emotion, typically six categories are used: joy, sadness, anger, fear, surprise and disgust. Ekman (P. Ekman 1999) claims that these six emotions, also known as basic emotions, are universally expressed in the same way, and that they are therefore archetypal. In general, categorical description is very clear-cut, and easy to communicate to actors. The disadvantage of this method is that it discards a great deal of information of the affective states: everyday affect is on a much wider spectrum than these six simple categories suggest.

Dimensional description describes an affective state along a set of main factors (dimensions) that can take continuous values. Commonly, the three dimensions that constitute the axes of the three-dimensional affective space are pleasure, arousal and dominance (Russell and Mehrabian 1977). Pleasure, or valence, expresses how positive or negative the affective state is. Arousal describes the level of activation. Dominance relates to the control the person experiencing the affective state has over the situation. Sometimes, the dimensional space is simplified to the two dimensions pleasure and arousal (Marsella, Gratch and Petta 2010) (Z., et al. 2009). The dimensional description allows us to represent each affective state with a certain combination of valence, arousal and/ or dominance on a continuous space. So where the categorical description is limited to a finite set of affective states, dimensional description can describe an infinite number of valence-arousal-dominance combinations.

2.1.2.2 Affect Measurement Tools

We discern two strategies for affect measurement: static and continuous. A static measurement represents an affective state at a certain point in time. A continuous measurement represents the progression of an affective state across time.

A traditional method for measuring affect is a questionnaire. Mehrabian and Russell created the Semantic Differential Scale (Mehrabian and Russell 1974), which derives scores on the three affective dimensions from a questionnaire consisting of adjectives that are rated along a 9-point scale. Similarly, the PANAS scale (Watson, Clark and Tellegen 1988) is a questionnaire used to measure general affective states. It uses two dimensions, Negative Affect and Positive Affect. The scales consist of a number of affective states, and the user fills in a number signifying to which extent they experience that state. Both scales can be used for static measurement. A disadvantage with these methods is that they are both extensive questionnaires, implying that it takes a great deal of time and effort to fill out and is troublesome for users who do not speak its language (Bradley and Lang 1994) (Soleymani, Pantic and Pun 2012).

The Self-Assessment Manikin (SAM) (Bradley and Lang 1994) is a measuring tool which uses pictures rather than words, overcoming language issues commonly associated with questionnaires. Originally created as an interactive computer program, the SAM was later expanded to be used simply with a paper and pencil as well. The SAM consists of three rows of figures (See Figure 2-1), each row representing one of the affective dimensions pleasure, arousal and dominance. While the use of pictures makes the tool much simpler to use than a questionnaire, it does require the user to understand the meaning of each dimension (Broekens and Brinkman 2013).

The AffectButton (Broekens and Brinkman 2013) is a tool with an emoticon face that can be changed by moving the mouse (Figure 2-2). The user selects the face that matches the affective state he/she has in mind, and presses the button. The AffectButton then provides feedback in the form of a Pleasure-Arousal-Dominance triplet. The button can provide static measurements in this way, and by adjusting the face while observing a stimulus the button can provide dynamic measurements as well. An advantage of the AffectButton over both questionnaires and the SAM is that it requires no understanding of a language or dimensions of affect; it simply requires knowledge of human facial expressions.

Trace tools are instruments that can measure a continuous representation of affect. The user is provided with a slider which represents one or more dimensions of affect, as depicted in Figure 2-3. A widely used trace tool for the annotation of stimuli is Feeltrace (Cowie, Douglas-Cowie and Savvidou 2000). The user watches a video or listens to a recording while moving a slider among two dimensions. A newer version of Feeltrace, GTrace, was released in 2011 (Roddy Cowie 2013). The user chooses a dimension (e.g., arousal) and moves the slider among that dimension while observing the stimulus. These trace tools can be used for trace annotation in a laboratory setting. Soleymani (Soleymani, Caro, et al. 2013) created a tool based on these two tools for online trace annotation of music.

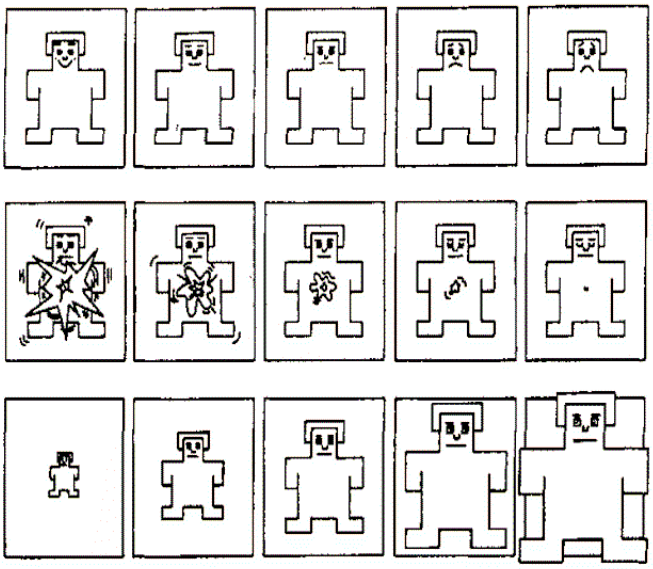
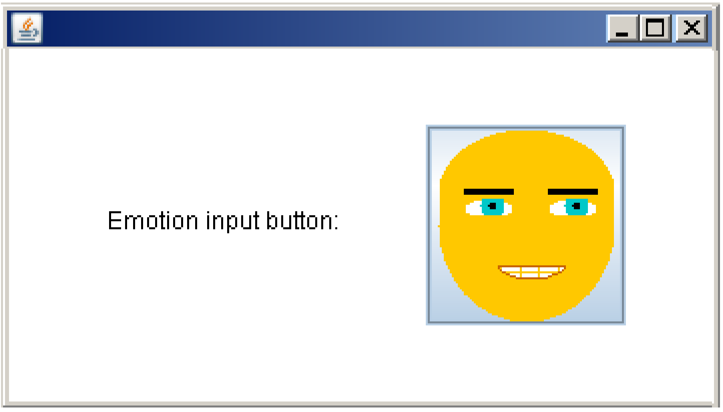


Figure 2-1: The SAM. The three rows represent Pleasure, Arousal and Dominance. The SAM is filled out by selecting one picture in each row that best represents the reported affective state.



Emotion input button:

Figure 2-2: The AffectButton. By moving the mouse, the user changes the face on the button. The user confirms by clicking the button.

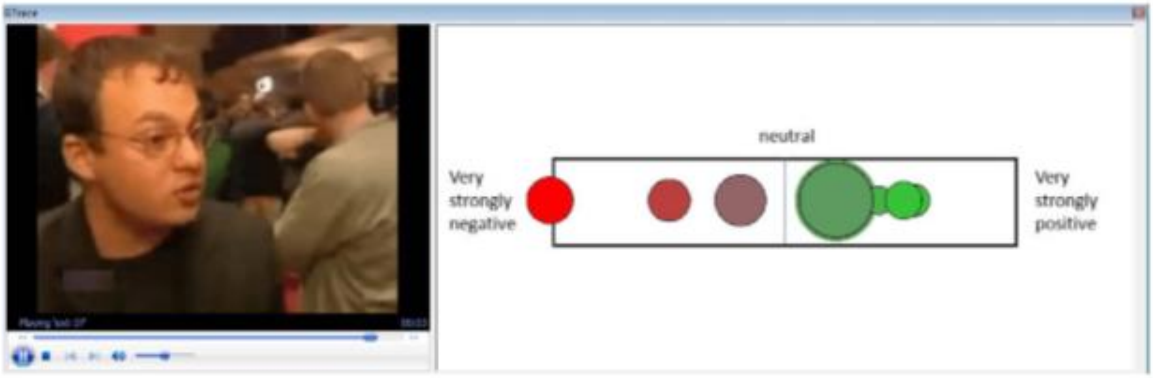


Figure 2-3: GTrace. The rating is performed by watching the video on the left and moving the slider on the right along the axis, leaving the coloured trace.

2.2 Affective Databases

Training an affect recognition system requires an annotated affective database: a collection of affective data and annotations (Z., et al. 2009). The database can contain several kinds of affective data, including speech, visual data (e.g. photos or videos), audio-visual data and physiological signals. This section, we cover the different aspects of the creation of an affective database. We will discuss how the affective content can be described, which design choices apply in the design of an affective database, and how such databases can be annotated. We will present examples of existing databases to illustrate how these choices are made in practice.

2.2.1 Designing an Affective Database

In this section, we will address the choices involved with the creation of the material for an affective database. Based on Douglas-Cowie (Douglas-Cowie, et al. 2003), we discuss three main issues: the scope, the naturalness and the context.

2.2.1.1 Scope

An obvious first choice to make is what to incorporate in the database. The scope of an affective database is defined by the range of affective states it covers, the subjects that express the affective states and the modalities the subjects use to communicate these states. Thus, a design for an affective database requires a choice in these three areas. Most importantly, the scope is determined by which affective states the database contains, as well as the nature of the affect. For example, one must choose whether to include emotions or moods, pervasive or episodic emotions (Cowie, Douglas-Cowie and Cox 2005).

Other choices to make are the number of different persons recorded, which languages are spoken and the gender of the subjects. For most affect recognition systems, it will be desirable to have the system learn from multiple people, to obtain a general model of affect that takes individual differences in expressing affect into account.

People communicate affect via multiple expressive channels, or *modalities*. Speech, facial expressions, gestures and body posture are the most notable examples of such modalities. To broaden the scope of an affective database, we can capture these different ways of communicating affect. For example, the FABO database (Gunes and Piccardi 2006) contains data recorded in two modalities. They used two cameras: one zoomed in on the face and one showing the whole body. The participants were asked to perform gestures using their face and body. The emphasis in this database was on expressions rather than feelings.

In addition, although in most databases, the data is recorded using a microphone and/or a camera, recently it is becoming possible to use more advanced input sensors. For example, it is possible to record the skeleton data of the subjects using the Microsoft Kinect. This creates a new modality separate from the different body parts. An example of a database that uses the Kinect is the FEEDB database (Szwoch 2013). It contains videos of 50 IT students showing posed facial expressions. The CALLAS database used Wii motion controls and data gloves in addition to two cameras in order to capture exact gesture data, allowing for new methods of multimodal affective analysis.

2.2.1.2 Naturalness

In the creation of an affective database, there is always a trade-off between control and realism (Sneddon, et al. 2012). This dilemma is most prominent in the choice of naturalness, or the choice of emotion expression. We can discern three different kinds of expression: natural, induced and portrayed (Bänziger 2007).

Natural expressions are expressions that occur in a real-life setting, typically making them the most spontaneous expressions. The researcher has no control over these expressions, and simply collects them. An example is a TV show or an interview. While these expressions are considered the most representative and authentic (Z., et al. 2009), they have a number of serious disadvantages. First of all, the researcher has no control over what kind of emotion is expressed: the emotions occur naturally. Secondly, it is very difficult to identify these emotions; it is difficult to establish exactly in what affective state the person was, and other information concerning the situation and context can be equally difficult to obtain (Bänziger 2007).The VAM database (Grimm 2008) contains natural emotional data. The videos were taken from German TV talk show “Vera am Mittag”, with the intention of obtaining emotions that occur naturally in a conversation rather than induced or posed emotions. The database consists of three parts: the videos, the audio and individual still frames of faces taken from the videos. The separated sentences in the audio were annotated using SAMs, with seventeen annotators for each audio fragment. The still frames were annotated using SAMs and emotional categorical annotations, with the number of annotators varying per image from 8 to 34, with an average of 13.9. The videos were not annotated.

Induced expressions are considered the midway between natural and portrayed affective states. The affective state is expressed in a controlled setting, the expression being a result of material selected by the researcher to intentionally induce a certain affective state. The advantage of this method is that the researcher has much more control than with natural expressions, while the resulting affective state is closer to spontaneous. These affective states are still hard to predict and control, however (Bänziger 2007).

Two examples of affective databases with induced affective states are The Belfast Induced database (Sneddon, et al. 2012) and the CALLAS database (Caridakis, et al. 2010). Each piece of data was created by giving the subject a task. Each task was made with the purpose of inducing a certain emotion. For example, the subjects were asked to place their hand into a box containing, unbeknownst to the subjects, cold spaghetti with sauce. This task was intended to induce disgust. The CALLAS database was inspired by the Velten mood induction technique (Velten 1968), in which a subject is put in a particular affective state by reading aloud a number of sentences. The technique will be explained in detail in Chapter 3. Using this induction method, the speech, gestures and facial expressions expressed by the subjects was recorded.

Finally, portrayed expressions are acted affective states. They are the most controlled type of expressions: the researcher can ask the actor which affective state to portray and how intensely. Since the actor can be instructed as the researcher wishes, this method makes it possible to record high quality data with multiple modalities. Of course, induction can be used to help the actors portray an affective state. There are some doubts, however, whether acted affective states are representative of natural affective states. Bänziger (Bänziger 2007) discusses some common objections against portrayed affective states. One of the main concerns is that portrayals are a reflection of stereotypes associated with affective states, rather than actual genuine affective states. Bänziger counters that actors have several methods at their disposal to elicit believable affective states, such as Stanislavski's techniques which activate the actor to re-experience emotions. See Cole (Cole 1995) for more details. Bänziger argues that using such techniques, believable affective states can be portrayed. She also proposes to use judgement studies to select which portrayals are credible.

The IEMOCAP database (Busso, et al. 2008) is a database of acted emotions. The database was created with the intention of incorporating interaction as much as possible into the pieces of data. They used dialogues and improvisations instead of monologues or short sentences. Each piece of data was a scene, based on a situation made beforehand. The goal was to have the lines expressed in a way they might be expressed in a real-life situation. The data was split in utterances, which were annotated individually using categorical descriptors and the SAM. The categorical descriptors were used by three annotators for each utterance, whereas the SAMs were used by two annotators.

2.2.1.3 Context

When humans communicate affect, how the affect is expressed and perceived depends heavily on the social context (Busso, et al. 2008). Therefore, many researchers argue that an affective database should include a social context for the affective states, rather than simply containing an isolated expression of an affective state. Such a context can be decided by the presence or absence of interaction, and the setting chosen for the recording of the data. We illustrate how context can be incorporated in an affective database with examples of existing work.

The IEMOCAP database (Busso, et al. 2008) was created with the intention of incorporating context as much as possible into the pieces of data. They used dialogues and improvisations instead of monologues or short sentences. Each piece of data was a scene, based on a situation made beforehand. This way, the lines said were expressed in a way they might be expressed in a real-life situation. Additionally, the interaction between actors also has an effect on the data. The approach of integrating context through the use of improvisation between actors was also used by the CreativeIT database (A. e. Metallinou 2010), which included scenes between two and eight minutes long and recorded with cameras, Motion Capture markers and close-up microphones.

The Belfast database (Sneddon, et al. 2012) contains induced emotions. Each piece of data was created by giving the subject a task. Each task was made with the intention of inducing a certain emotion. So while the IEMOCAP and CreativeIT databases incorporated context in the form of an imagined situation, with Belfast the context was the direct cause of the emotion, and a natural part of the data.

The SEMAINE database (McKeown, et al. 2012) creates a social context by having users interact with an artificial character created to engage in conversation with the user. The data is split in three groups: one where the user interacted with a character played by an operator, one where the character spoke in pre-defined phrases selected by the operator, and one where the character was fully played by an agent. Thus, the SEMAINE database shows natural emotions in a controlled social context.

2.2.2 Annotating an Affective Database

As stated earlier, an affective database requires annotations to describe its affective content (Z., et al. 2009). In an ideal case using actors, we would be able to fully control the affective state of the actor and provide that exact affective state as annotation. However, due to the subjective nature of affect, it is impossible to quantify precisely which affective state is finally contained in the acquired content. Therefore, it is common practice to rely on external observers to report the affective state they perceive in the acquired data (Douglas-Cowie, et al. 2003).

It is important that a robust methodology is chosen to annotate the data. First, the researcher needs to choose a model to describe the affective states expressed in the data. As we described in Section 2.1.2 , the two most commonly used models for affect recognition are the categorical and dimensional models (Z., et al. 2009). Then, an appropriate measurement tool (see Section 2.1.2.2) needs to be chosen to allow annotators to quantify affect in a reliable way, and such that it conveys information useful for research (e.g. a choice has to be made on whether the annotations should be continuous, static or both).

The HUMAINE database (Douglas-Cowie, et al. 2007) illustrates a wide variety of annotation methods. Created by selecting clips from different existing databases, it is a collection of six induced reaction and three naturalistic databases of varying size and modalities. Fifty clips have been extensively annotated statically in eight dimensions and continuously using eight kinds of traces. These dimensions include not only descriptions of the affective states, but also contextual descriptions such as key events, the physical setting, the social setting, the degree to which the person is attempting to hide their affective state. Thus, the annotation provides a more complete description of the affective state than simply expressing it as a number, taking the complexity and context-sensitivity of affective states (D'mello and Kory 2015) into consideration.

Annotation of affective data is often done with the use of expert annotators. However, the perception of affect is still a subjective matter, and so it is difficult to evaluate the quality of the resulting annotations (Siegert, Böck and Wendemuth 2014). Some annotators may be more compassionate or empathetic than others, resulting in a difference in perception of the affect (Devillers, Vidrascu and Lamel 2005). A commonly used method to overcome subjectivity is to combine the opinions from a multiple annotators (at least three), for example via majority voting (Vidrascu and Devillers 2005) or taking the mean value of all annotations (Grimm, Kroschel, et al. 2007). Even using this strategy, the annotations are only appropriate for use if they are *reliable*: if they can be assumed to have been performed in an accurate and consistent process, and thus can be used to train an accurate recognition system. A commonly accepting measure for reliability of annotations is the agreement on the assigned annotations between annotators: if different annotators produce consistently similar results, then it can be inferred that they will perform reliably (Artstein and Poesio 2008) (Gwet 2008). In conclusion, it is desirable to have *multiple* annotators to increase objectivity, and to have high *agreement* among these annotators.

2.3 Crowdsourcing

2.3.1 Basic functioning

Howe (Howe 2006) introduced the term crowdsourcing with the following definition:

“Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.” (Howe 2006)

In fact, crowdsourcing is based on the principle of the Wisdom of the Crowd (Brabham 2008). Empirical studies have shown that certain problems can be better solved by a large group of people than a small group of experts (Surowiecki 2004).

Crowdsourcing is widely used in both academic and industrial contexts, with varying goals. Gadiraju et al (Gadiraju, Kawase and Dietze 2014) categorized different crowdsourcing tasks in the following types, based on a survey on the crowdsourcing platform CrowdFlower:

- *Information Finding.* The worker is asked to search for information on the internet. For example, “Find information about a monument in Amsterdam”, or “Find the cheapest air fare for the selected dates and destinations”.
- *Verification and Validation.* The worker is asked to verify certain aspects or validate given content. For example, “Is this a Spam Bot? Check whether the twitter users are either real people or organisations, or merely spam twitter user profiles”, or “Match the names of personal computers and verify corresponding information”.
- *Interpretation and Analysis.* The worker is asked to process information. For example, “Categorize the age of persons in a photograph”, or “Provide affective annotations for a video”.
- *Content Creation.* The worker is asked to generate new content. For example, “Suggest a design for a new product”, or “Translate the following content into French”.
- *Surveys.* The worker is asked to fill out a provided questionnaire. For example, “Take part in a psychological survey from the University of Tilburg”.
- *Content Access.* The user is asked to merely access certain content. For example, “Watch the video provided in the link”, or “Click the link and read the information.”

In the domain of paid crowdsourcing, the person who outsources the task is typically called an employer, and the people who execute the task are called workers. For quick access to workers, an employer can submit his task to a crowdsourcing platform. A crowdsourcing platform is an online marketplace where employers can submit their tasks, and workers can browse and decide to perform these tasks. Three major crowdsourcing platforms are Mechanical Turk[1], Microworkers[2] and CrowdFlower[3]. After the worker has completed the task successfully, the employer pays the worker a small fee. Using such a platform, the task of recruiting workers can be completed more efficiently than recruiting an expert (Snow, et al. 2008). Tasks can be performed at much greater speeds while spending less money (Hsueh, Melville and Sindhwani 2009).

2.3.2 Crowdsourcing for affective annotation of media

Crowdsourcing offers many possibilities to the field of affective computing. Recent studies have emerged using crowdsourcing for the collection and annotation of affective media, as well as powering affective applications. See (Morris and McDuff 2014) for an overview. In this section, we will discuss the motivation for using crowdsourcing for affective annotations of media and review some examples of prior work using this method.

We mentioned the importance of using at least three expert annotations in Section 2.2.2 . This approach has two issues. Firstly, annotation by three experts is often a costly and time-consuming process, making it unfeasible for a large database (Devillers, Vidrascu and Lamel 2005). Secondly, it is argued that expertise in a field related to affect does not necessarily

[1] https://www.mturk.com
[2] https://microworkers.com/
[3] https://www.crowdflower.com/

imply expertise in the experience of affect (Cowie and Cornelius 2003). A cheaper and faster alternative to expert annotation is to rely on the emotional wisdom of the crowd. It is suggested that using crowdsourcing, we can acquire annotations with the same reliability as that of expert annotations (Snel, et al. 2012) (Tarasov, Cullen and Delany 2010). Recent works have examined the reliability of crowdsourced annotations compared to the results from work performed by experts. Snow et al (Snow, et al. 2008) examined the use of crowdsourcing for language annotation tasks. One of these tasks was affect recognition, where affective categorical annotations were collected. They selected previous experiments with experts and attempted to recreate the results using crowdsourcing. They calculated the performance of each expert annotator by training a classifier using the resulting annotations and calculating the average Pearson correlation between the resulting annotations and gold standard data acquired from other experts. They used the same process to calculate the performance for the non-expert annotators. They found that four non-expert annotators for each item can offer the same performance as one expert annotator for all items. In similar experiments, Hsueh (Hsueh, Melville and Sindhwani 2009) and Sheng (Sheng, Provost and Ipeirotis 2008) drew the same conclusion.

Nowak (Nowak 2010) annotated a dataset of images using both expert annotators and non-expert annotators acquired via Mechanical Turk. The dataset was annotated in 53 categories, such as illumination, location and time. Nowak measured the accuracy by examining how often the same image was annotated identically. He found high accuracy between non-expert annotators, close to the accuracy among expert annotators. Using majority vote to acquire the non-expert annotations, he compared the resulting free-marginal kappa statistic (Randolph 2005) with the kappa statistic of the expert annotations and found they were similarly high. He filtered out noise in his crowdsourced annotations by using majority vote, a common practice for crowdsourced affective annotations (Snow, et al. 2008) (Eickhoff and Vries 2013).

We will discuss a number of examples of crowdsourcing campaigns with the goal of annotating affective data. Riek et al (Riek, O’connor and Robinson 2011) used gamification to acquire discrete labels for the annotation of videos. In a game which ran on Facebook, the annotators were asked questions about a video and were granted points for giving an answer that most annotators gave. They calculated Krippendorff’s alpha (K. Krippendorff 1989) over the answers by all the annotators. Obtaining an alpha of .7, they concluded that the results were reliable. A game is an effective way to engage the user to the task, and user engagement is linked to the user’s motivation to do their best at the task (Hoßfeld, Hirth, et al. 2014).

Baveye et al (Y. Baveye, et al. 2013) created a database of video clips extracted from movies. They used CrowdFlower to annotate the clips by sorting them along the pleasure axis of the dimensional space. A participant was shown two clips, and asked which conveyed the most positive emotion. They based their decision to use paired comparison on the notion that agreement on emotion is higher when it is described in relative terms rather than in absolute terms (A. a. Metallinou 2013).

Adabi et al (Abadi, et al. 2014) annotated a database of movie scenes using both Mechanical Turk and CrowdFlower. They collected dynamic annotations using a trace tool, and recorded the participants using their webcams.

Soleymani et al. (Soleymani, Caro, et al. 2013) also used a trace application to obtain dynamic annotations. In this manner, they annotated a music database using Mechanical Turk. They saw higher agreement for arousal ratings than for valence ratings (Soleymani, Caro, et al. 2013). This is relevant for our research. Arousal is the key difference between the two moods we are interested in, as anxiety and depression are both signified by low pleasure but have opposite arousal values.

2.3.3 Reliability issues in crowdsourcing

While crowdsourcing is a cost-efficient way of getting tasks done, the employer has far less control over the setting than with a small amount of experts. Workers can lose interest and become careless in their work, and some rush through the tasks to maximize their income. It is also possible that workers misunderstand instructions, due to a poor task design, or because of language limitations (Hirth, Hoßfeld and Tran-Gia 2013). Considering these issues, an examination of the reliability of crowdsourced annotations is required. We will present how this reliability compares with that of experts and give an overview of techniques used to control the reliability.

There are different approaches for dealing with unreliable task completion. Firstly, it is important to set the task up in such a way to keep the user motivated, for example to phrase the questions in a non-repetitive way to prevent automation (Eickhoff and Vries 2013). Secondly, there are several ways to verify the reliability of the worker. This can be done either before the task with a test or a screening (Downs, et al. 2010), during the task with several verification mechanisms (Hoßfeld, Keimel, et al. 2014), or after the task during the analysis of the data (Hoßfeld, Hirth, et al. 2014). In this section, we will discuss several mechanisms developed with the goal of preventing or detecting unreliable worker behaviour.

In the affective annotation of a song database, Soleymani (Soleymani, Caro, et al. 2013) ran a pilot experiment, where workers were asked questions about two songs, demonstrating their understanding of the dimensional model and their willingness to put effort into annotating a song. Workers who answered sufficient questions correctly were invited to the main task. 36.9% of the initial participants were invited to the task. Thus their method eliminated a majority of workers from the campaign beforehand who failed to properly perform the relevant task.

Hoßfeld et al (Hoßfeld, Keimel, et al. 2014) specified a number of elements of a crowdsourcing task that should be taken into account to increase the reliability of the results. As this work focusses on tasks that rate the quality of experience, not all of these elements apply to every crowdsourcing task in general. However, a number of mechanisms are relevant in other contexts as well. The following reliability mechanisms can be incorporated into the task.

- *Verification tests.* Ask the user a question about common knowledge. For example, “Which is larger, a mouse or a train?”
- *Consistency tests.* Ask the user a question, and later ask the user a similar question to see if the answer remains consistent. For example, “In which country were you born?” and later, “In which continent were you born?”
- *Content questions.* In an annotation task, ask the user to describe the content in the media they had to annotate. For example, “How many people appeared in the video?”
- *Gold standard data.* Ask an objective question relevant to the task for which the answer is known beforehand (Hsueh, Melville and Sindhwani 2009). For example, comparing a descriptive (objective) annotation with trusted previous annotations.
- *Application-layer monitoring.* Keep track of the actions of the user and the time these actions took. For example, monitoring the amount of time it takes the user to submit the answer to each question.

It is important to limit the number of verification items and questions so as not to make the task overly lengthy and tedious.

Kittur (Kittur, Chi and Suh 2008) created a crowdsourcing task to rate the quality of Wikipedia articles. He increased the correlation of crowdsourced ratings with expert ratings by including objective, verifiable questions relevant to the task for the annotators. We can see this as the same category as content questions in Hoßfeld's terms. Another way to detect unreliable responses is by checking the submitted results for work with suspiciously shorter task durations than average. In this case, Kittur checked for work where the task was done in such a short time that the user was unlikely to have read the full article in that time. In other words, he used application-layer monitoring.

Hirth (Hirth, Hoßfeld and Tran-Gia 2013) evaluated two approaches to tackle the problem of untrustworthy workers by applying them to a variety of tasks on Mechanical Turk and Microworkers. The two models were majority decision and a control group. Using the majority decision mechanism, the employer considers the results agreed upon by most workers to be the ground truth. Assuming the majority of the workers are reliable workers, the unreliable work will be filtered out. When using a control group, the employer first lets a group of workers perform the desired task and then recruits a second group of workers to check the results of the first task. While both models are useful for objective, simple tasks, it is more difficult to decide on or check the result of a subjective task.

2.4 Conclusion

We have established the theoretical differences between mood and emotion and compared different methods for the representation of affect. We will discuss how these principles influenced our practical decisions in Chapter 3. We have identified key choices that need to be made in the design of an affective database, from the scope, naturalness and context to the manner of annotation. Having established an overview of the considerations that are to be taken into account with regards to these design issues, we will explain our choices based on these considerations in Chapter 3, and discuss the process of creating the database in Chapter 4.

Finally, we have discussed crowdsourcing as a tool for the annotation of an affective database. Using a crowdsourcing platform to connect to a large amount of workers, the large task of annotating a database can be outsourced and performed in small parts by these workers. While this is a promising alternative to the traditional method of hiring experts to perform the annotation, it is vital that the task is designed in such a way that unreliable worker behaviour can be detected. We have explored a number of ways to deal with this behaviour and ensure the reliability of the task results. Our choices will be discussed in Chapter 3, and the crowdsourced annotation process will be explained in-depth in Chapter 5.

Chapter 3: Methodology

Having delved into the literature, we now present the methodology we will follow to answer the research questions outlined in the introduction. As stated in Chapter 1, our goal is to create a database of videos portraying people in a non-interactive context, showing anxiety and depression so that they represent the mood of elderly in care centres. We will now discuss how we plan to create this mood database and answer our research questions.

The remainder of this chapter is organized as follows. First, we discuss our decisions regarding the creation of the database, followed by our choices concerning the annotation of the database. Finally, we describe how performing these steps would allow us to answer our research questions.

3.1 Database Material Creation

This section describes the design and creation of the material that makes up the database. As the data is to be used for the recognition of moods in a non-interactive setting, there is unlikely to be any speech. Thus, we focus on visual data. As mood is a subtle, long affective state that takes time to develop after the cause, the database needs to portray long displays of affect that gradually develop. Therefore, we chose to create a video database.

The two main moods that the recognition system will need to recognize are anxiety and depression. Therefore, the system must be trained using visual data depicting these two moods. Additionally, we include training data depicting a counterexample: a display of mood that would be very different to the main moods, to show as much contrast as possible. Anxiety and depression are both negative moods (low pleasure), so we decided to use amusement as a counterexample, as it is a positive mood (high pleasure), in order to maximize the contrast.

As discussed in Section 2.2.1 , the creation of the material for an affective database is dependent on three issues: the scope, the naturalness and the context. We will describe our design according to these factors.

3.1.1 Scope

The *scope* of our database consists of three moods. Anxiety and depression are the two main moods. In addition, we included the mood amusement, which serves as a counterexample to the main moods for the mood recognition system.

Affect is displayed via multiple expressive channels, or modalities. It is claimed that information from multiple modalities is combined for human affect recognition (Gunes, Piccardi and Pantic, From the lab to the real world: Affect recognition using multiple cues and modalities 2008). Specifically, it is suggested that the most significant modalities are the facial expressions and body gestures (Gunes and Piccardi 2009). We decided to incorporate these two channels in our database. This was done using two cameras for recording, one of which was zoomed in on the actor's face and one which recorded the whole body. In addition, we recorded 3D skeleton data using a Kinect system, to provide additional features for postural analysis. As the setting for our videos is non-interactive, audio was not included.

3.1.2 Naturalness

In Section 2.2.1.2 , three kinds of affect portrayal were distinguished in terms of *naturalness*: acted, induced and natural. For a natural database of mood in elderly, we could hypothetically record elderly residents in a care centre. However, this comes with two important problems. Firstly, there are obvious ethical issues with recording elderly in their daily lives, especially when they are feeling anxious or depressed. Secondly, it is difficult to verify what exactly the mood was they experienced; the observed person might not even be aware of the mood they are experiencing. They will at least have trouble describing which mood they experienced, not to mention quantify the intensity of the mood. (Douglas-Cowie, et al. 2003) (Bänziger 2007)

Using induction, we would have more control over the situation. We would induce the moods we desire into the subjects, and record the subjects while in this mood (Sneddon, et al. 2012). While induction alone would allow us to gather the desired visual material, considering the purpose of capturing a long affective state that affects a person's cognition and behaviour, the advantages of using actors are considerable.

An acted database would give us the most control: we can ask an actor to perform the kind of behaviour we seek in any controlled setting. Actors can act in a personal and believable manner in an imagined situation. They use their body as a tool for expressing affective states (Gross, Crane and Fredrickson 2010), allowing us to capture subtle expressions. Furthermore, as the setting in which we record the data is fully controlled, we can choose the area, the lighting, the background and other factors that affect the quality of the resulting recordings. A disadvantage of this option is that the resulting material could be less natural than the other two options; it is disputable whether scripted behaviour of a depressed elderly can be close enough to real behaviour to be useful for a database. (Batliner, et al. 2004)

Thus, we decided to combine induction and acting. We induced the desired mood in the actor, and then recorded the actor while he performed actions relevant to the database. To ensure the footage would not simply show an actor following instructions, we chose to work with improvisation. Improvisation has been shown to evoke more spontaneous affective states than scripted scenes. (Busso and Narayanan 2008)

3.1.3 Context

In order to get a realistic recording, a clear *context* for the moods was needed. In particular, we required appropriate scenarios and instructions for the actors. The database is to be used to train a system that can recognize negative moods of elderly in a care centre in the same way a caretaker would. Therefore, we needed a closer understanding of which behavioural cues caretakers capture to infer the presence of a negative mood. We decided to interview caretakers at a care centre for this purpose.

Our context is defined by the requirements of the database. The actors play out a number of scenarios which were based on the interviews to create a situation which could occur in a care centre and cause the mood in question. They are set in the room of an elderly who lives in a care home. Thus, the person has no interaction with others.

3.2 Crowdsourced Annotation

In this section, we describe how we annotated the database. The goal was to acquire reliable annotations, properly describing the mood(s) portrayed by the person in each video. The annotation phase will be explained in detail in Chapter 5.

We have mentioned that, due to the unfocussed and personal nature of mood (Beedie, Terry and Lane 2005), mood recognition is more difficult than emotion recognition for humans. Previous work (P. Ekman 1999) has stated that the recognition of emotion and the recognition of mood are related. Therefore, we expected it to be useful for future work (Katsimerou, Redi and Heynderickx 2014) to have the database annotated in terms of both the mood and emotions that were portrayed by the people in the videos. We aimed to have the database annotated in terms of the general mood of the videos and progression of emotions among time. In other words, we used static annotations for mood and continuous annotations for emotions (we discussed these two kinds of annotations in Section 2.1.2.2).

We decided to use crowdsourcing to acquire these annotations. As a basis for our annotation tool, we used the tool proposed by Soleymani (Soleymani, Caro, et al. 2013) for the affective annotation of music. This tool was based on the trace tool known as GTrace (Cowie and Sawey 2011) (see Section 2.1.2.2 for a review of these annotation tools). We adapted Solyemani's tool implementing three new features:

(1) the audio player was replaced with a video one;
(2) the annotation toolset was expanded to include the Self-Assessment Manikin and the Affect Button, to allow the static annotation of mood and
(3) a number of reliability control mechanisms were incorporated throughout the annotation task. The latter were especially important towards addressing one of the core concerns in this thesis, which is related to the extent to which crowdsourcing can be used to collect high-quality affective annotations.

These mechanisms, along with the detailed annotation protocol defined to collect the data, will be discussed in-depth in Chapter 5.

3.3 Analysis

Our analysis of the results consisted of two parts: investigating to what extent the videos showed the mood we intended to portray and analysing the reliability of the crowdsourced annotations.

Firstly, we were interested in the difference between the mood we intended to induce and the mood an observer perceived. For every video we made, we had a recording of each of these moods. We compared the mood we intended to induce (anxiety, depression or amusement) with the SAM value given by the annotators.

The second part concerned the issue of reliability, also discussed in Chapter 2. As stated in the previous section about the annotation, we implemented a series of reliability control mechanisms. Each of these mechanisms provided a filtering criterion. We made an analysis of how employing these mechanisms changed the reliability of the annotations remaining after filtering according to the criteria. Following common practice (Nowak 2010) (Calvo, et al. 2014), we measured reliability in terms of agreement between annotators. We calculated Krippendorff's alpha (K. Krippendorff 2011) as a measure of the inter-rater agreement after each filtering stage. We performed this analysis to gain insight in the effectiveness of each mechanism, to examine the reliability of the resulting annotations. The analysis and its results will be discussed in Chapter 6.

Chapter 4: Database Creation

A database can consist of acted, induced and/or naturalistic affect. For the purpose of training a system that can be applied in practice to recognize spontaneous moods, it is desirable to provide training videos that are as spontaneous as possible. At the same time, it is also desirable to maintain control over the moods that are shown. We decided to use actors, who specialize in using their body and facial expression to show their affective state. To make the moods more spontaneous, we decided to induce the desired moods in the actors. Thus, we combine the actors' expertise in expressing affect with the naturalness of induced affect as opposed to portrayed affect.

As emotion is a direct reaction to an object, a single stimulus is often sufficient for emotion induction. Mood regulation is a more complex matter. (W. Morris 1992) (Parkinson, et al. 1996) (Thayer 1996) Moods are longer and have no specific cause or direction. (Lane 2000) Therefore, unlike many induction methods, we should not use a single stimulus to evoke the mood. We should look for a mood induction method that generally changes the subject's mood without a specific cause. Mood changes slowly and continues to linger. Therefore, induction should take some time to let the mood sink in. After considering different state-of-the-art induction procedures (Westermann, Stahl and Hesse 1996), we found music to be the most appropriate induction method (Pignatiello 1986).

We used the literature to decide how to instruct our actors. It has been proposed that moods are easier to hide than emotions, and more personal. Furthermore, moods are less intense than emotions. Thus, we seek a subtle display of affect. Some researchers, such as Parkinson (Parkinson, et al. 1996), have suggested that moods are expressed via bodily posture. Therefore, we were interested in the physical display of mood.

It is widely accepted that moods are connected to thought processes (Beedie, Terry and Lane 2005). Whereas emotion evokes physical reactions, mood is affected by, and affects cognition (Davidson 1994). This means that when designing a mood database, the context, situation and thoughts of the subject are very relevant to the mood itself. Specifically, in an acted mood database, instructions for the actor should include details on what the character is thinking of while in the specified mood. In other words, a clear *scenario* with a character must be specified. Furthermore, while emotions are spontaneous reactions, a mood is a reaction to a cumulative sequence of minor incidents (Beedie, Terry and Lane 2005). This is a fact that these scenarios must incorporate. We seek to record affect that is a result of a number of events that either took place prior to the recording or take place during the recording, causing the mood. Moods endure longer than emotions. Additionally, moods change more slowly and continue to linger. This means that the scenarios must be long, with no sudden large changes in moods.

As mood is governed by the brain, we required the actor to consciously affect to his or her mood. It is argued that it is difficult for someone in a given mood to indicate what the cause of that mood is. Russell and Feldman-Barrett (Russell and Feldman-Barrett 1999) stated that a mood usually has many causes, and that some of those can be difficult to detect. This means that, to evoke a specific display of mood, we should not be focussing on a specific cause of the mood, but rather their trail of thought that has been biased by the mood (Davidson 1994). Additionally, moods have been described as vague and unfocussed, and less defined than emotions. It decided not to instruct the subject to display a certain mood, but instead to provide him or her with the situation and induction that causes the mood.

To make our scenarios as realistic as possible, we first investigated the context of our videos. As our goal is to teach a computer to recognize negative moods of elderly, we first investigated how human experts recognize these moods. For this purpose, we conducted interviews among caretakers in care homes to find out what circumstances cause the two negative moods, and how elderly people express these moods. We used these results as a basis for the scenarios and instructions for the actors.

Each video shows an actor in one of the three moods. For each video, we induced the desired mood in the actor. Then, we recorded the actor while portraying a scenario created from the results of the interviews.

In this chapter, we will describe the process followed to obtain the scenarios from interviews with caretakers, and how these were combined with induction methods to obtain acted yet naturalistic portrayals of affect. Then we will explain how we used our instructions and our chosen induction method to create the recordings.

4.1 Gathering requirements: Interview with Caretakers

The interviews had two main goals: to identify a suitable context for the scenarios for the actors, and to find out which physical cues were used by caretakers to recognize these moods. We define physical cues as behaviour that can be observed from a person, such as a movement, gesture or facial expression.

As the research was focussed on the outward display of mood, we decided to interview the caretakers, who have to recognize their moods and act accordingly as a part of their job. In short, they are experts at elderly mood recognition. Because of the subjective nature of mood recognition, we felt it would be best to use semi-structured interviews. In this manner, we could ask for examples, insist on important details and investigate the nature of mood in depth.

4.1.1 Procedure

For the interviews, we contacted the Pieter van Foreest[4] care organisation. We visited three locations in Delft and The Hague. Two of these locations were meeting centres where elderly with dementia meet for daily activities, and one was a care home where elderly with dementia live. We interviewed six caretakers in the meeting centres and three in the care home. Due to limited availability in the care home, the three caretakers requested their interview to be taken with all three caretakers at once. All nine caretakers were female. The interviews lasted 45 minutes on average. Each caretaker was informed that she could leave at any given time.

The interviews were semi-structured: we had a list of topics and questions to discuss, but the flow of the interview could differ depending on the responses by the caretaker. We asked the caretakers to describe the way the elderly expressed mood, but also to give a demonstration if they were willing, which we recorded with a Sony HD Camera. All questions about mood were asked twice: once about anxiety, and once about depression. The interviews were recorded with an iPad, for which the caretaker gave her consent. After the interview, the caretaker was compensated with a “VVV cadeaubon” coupon worth 10 euros.

Our questions fell into six categories:

- **Explorative questions.** These questions were meant to make the subject of mood recognition clear to the caretaker. To start off the interview, we asked general

[4] http://www.pietervanforeest.nl/

questions that were easier to answer than more in-depth questions. An example of an explorative question is “How can you tell someone’s mood?”

- **Mood Assessment.** These were specific questions meant to gain an in-depth understanding of the way caretakers recognize mood. Specifically, we asked about what they see physically in each mood. For example, “If the resident is feeling anxious, how can you tell?”
- **Mood Assessment over time.** These questions discussed how knowing the elderly changes how the caretaker perceives their mood, and how much time it takes to learn the characteristics of their mood. For example, “Does learning the personality of a person help in assessing their mood?”
- **Context.** The goal of these questions was to learn what circumstances can affect the elderly’s mood. Recreating these circumstances could help the actors use a believable context to influence their mood. An example of a question in this category is “What might give a resident cause for sadness?”
- **Activities.** These questions investigated what a resident in a care home does in his or her room, and how much time he or she spends there. We wanted to get a clear picture of the daily life of residents in a care home, as they would be the characters portrayed by the actors. An example is, “What are the typical activities of a resident in his/her room?”
- **Ambience.** These questions discussed how the ambience, such as weather and lighting, can affect an elderly’s mood. For example, “Is the residents’ mood influenced by the weather?”

The full list of questions can be viewed in Appendix A.

4.1.2 Interview Results

We transcribed each interview by listening to the recorded audio and writing down the caretaker’s responses under the question that best described the subject. We then reported which answers were given multiple times and tried to identify common themes in each category. The following results were found, grouped per category:

- **Explorative questions.** The caretakers usually determine a person’s mood by their attitude and posture. For example, a closed posture (e.g. folded arms) can signal a negative mood whereas an open posture with a lot of gestures can be a sign of a cheerful mood. They also compare the person’s behaviour to the behaviour they are used to seeing from that person. This is in agreement with claims in the literature that recognition of mood is mainly done by recognizing bodily displays. (Parkinson, et al. 1996)
- **Mood Assessment.** For sadness, a common theme was passivity: a lack of display of interest in other people or activities. Notable signs of anxiety were constantly looking around and performing a lot of actions, often not relevant to the situation at hand. Additionally, a number of physical signs, bodily and facial, were named, which we list below.
- **Mood Assessment over time.** The most important thing when meeting a new elderly is to establish a baseline; the caretaker tries to discover what the person’s typical behaviour and postures are in a neutral mood.
- **Context.** Two causes for both anxiety and sadness were named: being reminded of negative experiences in the past, and being confronted with the fact that they are starting to lose grip on their lives due to old age and dementia. Additionally, a common cause for anxiety was a divergence from their daily routine, such Christmas days with special activities and moved furniture for a feast.

- **Activities.** Most residents spend their time in the living room. Activities in the room include watching television or having a nap. When anxious, some residents disarrange their room, for example by emptying their closets on the floor or moving their beds.
- **Ambience.** Caretakers largely agree on this topic: rain makes the elderly more sombre while abundance of daylight brightens their moods. As the day goes on, visitors can get tired, worsening their moods. More intense light makes them more active. Too many stimuli can cause distress.

4.1.2.1 Cues

We used the answers to the mood assessment questions to get a list of physical cues that caretakers observe when an elder is in anxious mood. Since these cues appear in a real-life situation, it is desirable to have these cues appear in the database. Additionally, as the intent is to use physical acting, these cues can be given to the actor as instructions in order to make them feel the intended mood, allowing them to portray that mood more realistically. We list the resulting cues below, sorted by mood and by part of the body.

Anxiety

Behaviour	Face	Eyes	Hands	Walking	Breathing	Sounds
• Constantly looking for something • Walking around the room • Going to the toilet often • Watching the clock a lot • Sitting on the tip of one's chair • Constantly trying to get up	• Tense muscles, a strained posture, tight lips, face pulled tight • Grinding of teeth	• Eyes focussed on one particular thing at a time • Eyes wide • Eyes looking for something: for confirmation, for safety, for something to grab hold of	• Moving one's hands a lot • Plucking one's clothes • Pulling or fumbling with the table cloth • Balled fists • Tapping one's hand on the table	• Walking crookedly • Walking unstably • Walking with the upper body more to the front than the knees. • Tapping on the floor with a foot	• Heavy • High in the chest • Fast	• Making a tutting sound • Tapping on the table • Coughing

(a)

Depression

Behaviour	Face	Eyes	Posture	Upper body	Breathing	Sounds
• Unwilling to perform any action • Reacting negatively to anything • Tired, falling asleep	• Absent-minded, expressionles s • A mistrusting, reserved expression • A hanging mouth with the corners turned downwards, No smile • Frowning	• Dull eyes, unfocussed • Avoiding eye contact • Weepy eyes	• Making oneself small, sitting slumped • Focus is turned inwards • Any gestures are done with very little force behind them; half-gestures	• Hanging elbows • Shoulders hanging downwards • Shoulders forward and arms towards each other	• Slow • Calm	• Sighing • Moaning

(b)

Table 4-1 Cues for Anxiety (a) and Depression (b) based on the interviews with the caretakers, sorted by part of the body.

4.1.2.2 Scenarios

We used the answers to the Context, Activities and Ambience questions to understand what the daily activities of an elder in a care centre are, and which situations can typically be a cause for anxiety and depression. From this, we devised a list of scenarios. By a scenario we mean a description of a situation in which an elder finds him/herself, causing a certain mood. We wrote these scenarios in a format that would allow us to explain them to the actors. The actors could play these scenarios to allow them to experience the mood more naturally. In addition, the scenarios let us provide the context that would also cause a negative mood in practice, in a care home. We selected the scenarios that were most useful for the recordings, resulting in two scenarios for anxiety and depression and one for amusement. An example scenario is given below; the full set of the used scenarios can be found in Appendix B.

Scenario Depression

Situation

The character is in his room. Each week, his son pays him a visit. He will be coming this afternoon. The character gets a phone call from his son. At the end of the chat, the son tells him he won’t be able to visit this week.

Instructions

The actor is sitting in his chair. The phone is on a desk. The actor picks up the phone and mainly listens. He plays out his son’s dialogue in his head, and he himself has four lines of dialogue: “Hello. Yes. Oh. Yes.” This conversation may take as long as the actor wishes. After the phone call ends, the actor may stand up, walk around, or sit as he pleases.

Intended moods

We expect a transition from happy to depressed. The character is happy at first, looking forward to his son’s dropping by. When his son calls, he is happy to speak to him again. The news at the end of the phone call makes him sad.

4.2 Mood elicitation in actors

We combined the controlled setting of using actors with the naturalness of using induction. The data from the interviews were analysed, giving us a list of physical cues and a list of scenarios, both of which we wanted to see in the videos of our database. As we concluded from the literature that we would focus on the physical display of moods, we decided to use physical theatre as an acting method. This section discusses our choices for methods in these two areas.

4.2.1 Mood Induction

Several techniques exist for inducing mood. These mood induction procedures (MIPs) are traditionally used to research the psychological effects of mood (Westermann, Stahl and Hesse 1996). The challenge for this work, on the other hand, was to select one or more MIPs to induce the actors to improve their performance.

Westermann (Westermann, Stahl and Hesse 1996) compared different MIPs in terms of validity and effectiveness. Out of these, he found the following MIPs to be most efficient for inducing positive or negative moods:

- **Imagination.** The subject imagines situations or events that induce a specific mood. This is a relevant MIP, as showing affect by experiencing an imagined situation is the core of acting.
- **Velten.** The subject reads aloud sentences with a certain affective weight. (E 1968) The Velten method has been criticized for demand characteristics (Buchwald, Strack and Coyne 1981). On the other hand, the Velten MIP was a promising possibility to use on actors, as they are used to reading aloud texts and feeling the affect behind it: they do the same thing when performing a monologue.
- **Film and story.** The subject watches a film or another form of narrative and the story elicits a certain mood. Gross and Levenson (Gross and Levenson 1995) found that film is indeed an effective method to induce affect, and for each of the basic emotions they created a list of films that are most effective for inducing that emotion. A disadvantage for the actors is that they are induced passively, as opposed to the Velten MIP, which allows them to use more of their capabilities as actors.
- **Music.** The subject listens to a piece of music that suggests a specific mood (Pignatiello 1986). While the pieces of music are generally selected beforehand, there are examples of studies, such as (Sutherland, Newman and Rachman 1982), where the subjects were asked to select a piece of music that they thought was most suitable to induce a certain mood to them. The advantage of the music MIP for actors over the other methods is that it can be used to enhance an actor’s performance during the performance itself. The mood can actively be changed during the performance with a change in music.
- **Feedback.** The subject receives feedback on his or her performance in a task. Usually, this feedback is false and specifically chosen to induce an intended mood. In our study, the director could tell the actor that he is not doing well in order to induce a negative mood. But this may create unwanted friction in the interaction with the actor.
- **Social Interaction.** The subject engages in pre-arranged social interaction. The subject takes over the mood of the person they interact with, or feel better after ‘helping’ that person. Actors typically use their interaction with their fellow actors to induce the emotions and mood for themselves. This is not applicable in our study, however, as the actors will be playing a role that is alone in his or her room.

Westermann found the film and story MIP to be the single most effective MIP for inducing positive moods, while all of these MIPs are approximately equally effective for the induction of negative moods. He mentions that MIPs can be combined, and recommends combining MIPs that can be used simultaneously.

The use of actors in scenarios implied the use of imagination induction. Considering the advantages of using music induction simultaneously with the imagination induction, we decided to use these two MIPs for the negative moods. Both during the induction and during the recording of the acting, we played music selected from research by (Albersnagel 1988) and (Västfjäll 2002). For anxiety, we used Stravinsky’s ‘The Rite of Spring’. For depression, we used Sibelius’ ‘Swan of Tuonela’ and a fragment from Dvoiak’s ‘Ninth Symphony’. For the imagination induction, we asked the actors to imagine they were experiencing the scenarios we deduced from the interviews. This process will be discussed in Section 4.2 . For the positive mood, a counter-example, we used film and story induction, following Westermann’s recommendation.

4.2.2 Acting techniques and setting

As the physical cues we had gathered had been described as a sign of anxiety or depression, we wanted these cues to show in our database as well. To make these cues more natural to the actor, we used a number of techniques from a style of acting known as physical theatre. The principle is that, by changing an external characteristic of his body, an actor can change his inner state of being (Stanislavski 1989). The techniques apply the James-Lange theory of emotions, which in its simplest form states that emotions are a result of physiological changes in our bodies (Cannon 1987): in other words, we do not cry because we are sad, but we are sad because we cry. Thus, theoretically, an actor could make himself sad by pushing his muscles to cry.

In our case, we decided to have our actors adopt the physical cues belonging to the mood we wanted to induce and see how this affected their mood. We used physical cues extracted from the interviews at the care centres. The actors could use the technique to experience an intended mood by adopting the physique associated with the mood (Stanislavski 1989). After giving the instructions, we gave the actors some time alone to 'get in the mood': they could find out how to walk around with the given physical cues, and find out what the physique and the music made them feel.

Once the actors had found their physique, we asked them to place themselves in the situation given in the scenarios. The scenarios left some room for imagination, so that the actors could fully immerse themselves in the situation. During the recording, we told the actor the description given under "Situation", and gave them the instructions described in "Instructions". They were instead not instructed directly on the specific mood they had to portray, to avoid the influence of demand characteristics and increase the naturalness of the recordings.

4.3 Recording

The next step was to decide the structure of our database, and the experimental setup of the recording of the videos. This section explains the decisions made in this design step.

4.3.1 Setup and environment

We sent invites to actors in our acting network, who were non-professional actors who had participated in advanced acting lessons. We scheduled fifteen actors for recording. Out of these actors, nine were female and six were male. Their ages varied from 15 to 28 years old.

We set up a laboratory room in the EWI faculty of Delft University of Technology for the recordings. The room was lit with LED ceiling lighting. The set consisted of a chair or couch, where the actor sat at the start of each video. We used a green blanket as a background to increase the precision of the Kinect (Wang 2013) (Ye, et al. 2012).

The scenes were recorded with two HD cameras and a Kinect sensor. The cameras recorded at a frame rate of 25 frames per second and a resolution of 1920 x 1080. One camera recorded the whole set and the whole body of the actor, while the other was zoomed in on the actor's face. The Kinect sensor was also set to capture the whole set and body, and was used to allow easy extraction of the body's skeleton data. The Kinect data was recorded using the KinectStudio software (Jana 2012). This software allows the user to inject the recorded data into any Kinect program, which treats it as if it were captured at the moment of playback.

4.3.2 Method

We first informed the actor of the purpose of the project. We explained that the recordings would be used for a computer system that can recognize mood, aimed at the elderly in care centres. We then asked the actor to sign a consent form and fill out a self-assessment form, to measure the actor's initial mood (as baseline for future analysis) using a Self-Assessment Manikin (SAM) (Bradley and Lang 1994).

We started with a warming up. The warming up was meant to loosen the muscles and prepare the actor to work with his body, as we would work by giving physical assignments. We then started the recordings, which covered three different moods. For each mood, we used the following procedure. We briefly explained the first scenario to the actor. We then performed the mood induction, which we discussed in the previous section.

After the mood induction, we asked the actor to fill out a second self-assessment form, to judge how well the induction worked. We then started the first shooting. We let the music play during the scene, and let the actor play his part without interference. Afterwards, if the actors indicated they had not really felt the mood or were not content with their performance, we gave some tips, for example by naming some of the physical cues that the actor could emphasize on more, and recorded the same scenario again. We continued a recording for as long as we felt that the actor was still using different ways to express his or her mood. In this way, we aimed to get recordings that were as rich in expression as possible, and to get a variety of video lengths. After finishing the recording for this mood, we asked the actor to fill in another self- assessment form, concluding this particular mood. We made a short break of 10-15 minutes, to allow the actor return to his/her baseline mood. After the short break, we repeated the procedure for the next mood. Figure 4-1 shows the flowchart of the procedure per mood.

The final mood was amusement, meant to give counterexamples to the two moods we wanted to research. As stated earlier, we used film and story mood induction. We asked the actors to read a comic they found funny, or to watch a comedian of their choice. We recorded them watching or reading.

Some actors were unable to perform all scenarios due to limited time availability.

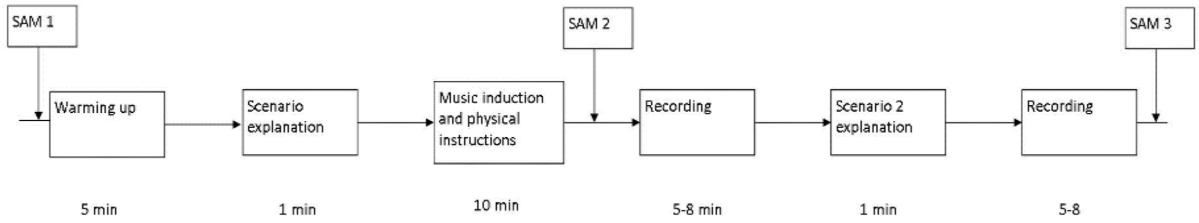


Figure 4-1: Flowchart of the recording procedure. This procedure was done twice: once for anxiety, and once for depression.

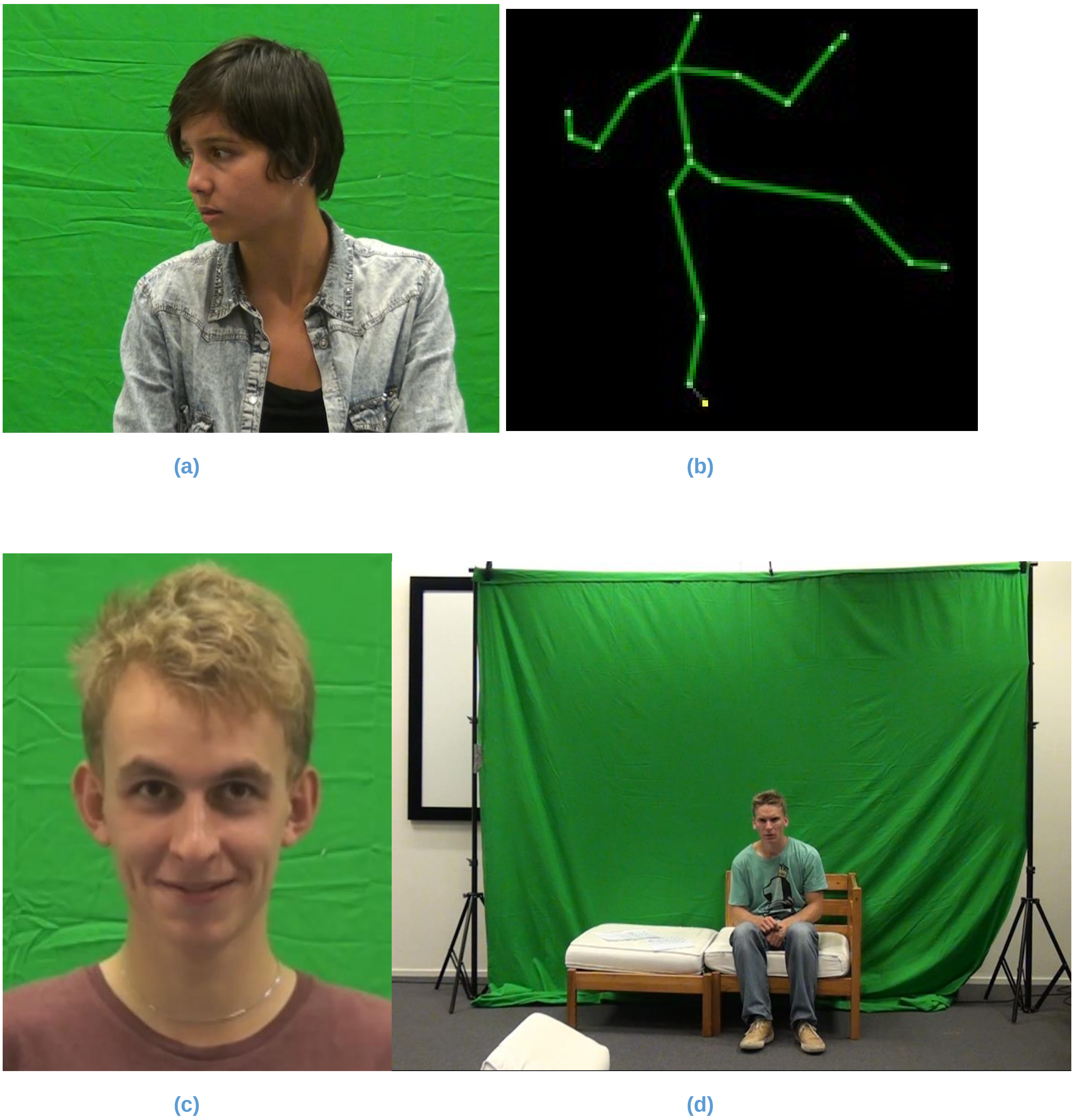
4.4 Results

In this section, we present the videos acquired in the recording phase and our analysis of the results from the self-assessment reports filled in by the actors.

4.4.1 Videos

The final recordings consist of 35 anxiety videos, 28 depression videos and 13 amusement videos, totalling in 76 videos with a total length of eight hours and five minutes for each modality. The shortest video has a length of 1:05 while the longest video has a length of 13:28.

We present an inventory of the gestures performed during the recordings in Table 4-2.

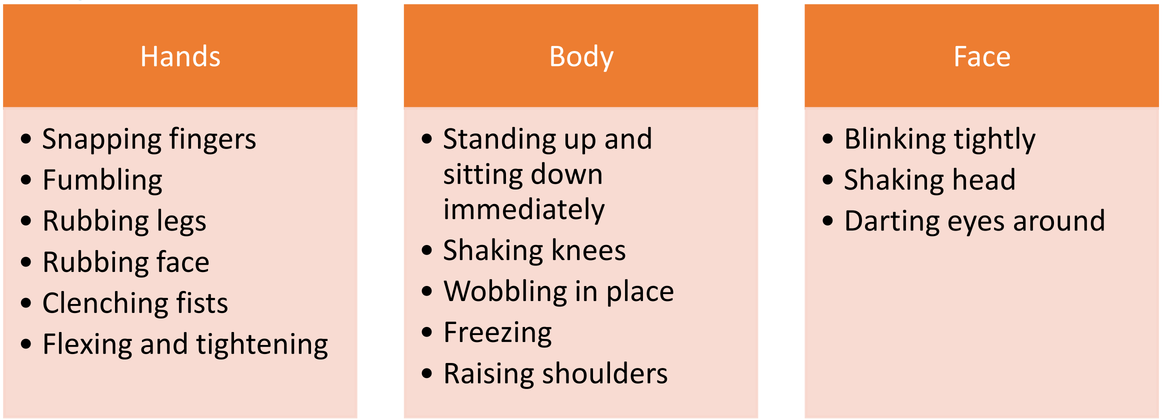


(a) (b)

(c) (d)

Figure 4-2: Snapshots from the resulting videos, taken from each kind of intended mood and used camera: from (a) the face camera with depression as the intended mood, (b) the Kinect with anxiety, (c) the face camera with amusement and (d) the body camera with anxiety.

Anxiety



Hands

- Snapping fingers
- Fumbling
- Rubbing legs
- Rubbing face
- Clenching fists
- Flexing and tightening

Body

- Standing up and sitting down immediately
- Shaking knees
- Wobbling in place
- Freezing
- Raising shoulders

Face

- Blinking tightly
- Shaking head
- Darting eyes around

(a)

Depression

Hands	Body	Face
• Rubbing eyes	• Casting head downwards	• Closing eyes • Casting eyes downwards • Blinking quickly • Sighing

(b)

Amusement

Hands	Body	Face
• Moving hand to cheek • Moving hand to mouth • Waving hand • Shrugging • Lightly touching fingers	• Sitting back • Moving body back and forth • Bobbing head from left to right • Cocking head backwards	• Laughing • Tightening eyes • Nodding

(c)

Table 4-2: Inventory of gestures in the recordings for Anxiety (a), Depression (b) and Amusement (c), sorted by the intended mood and the type of gesture.

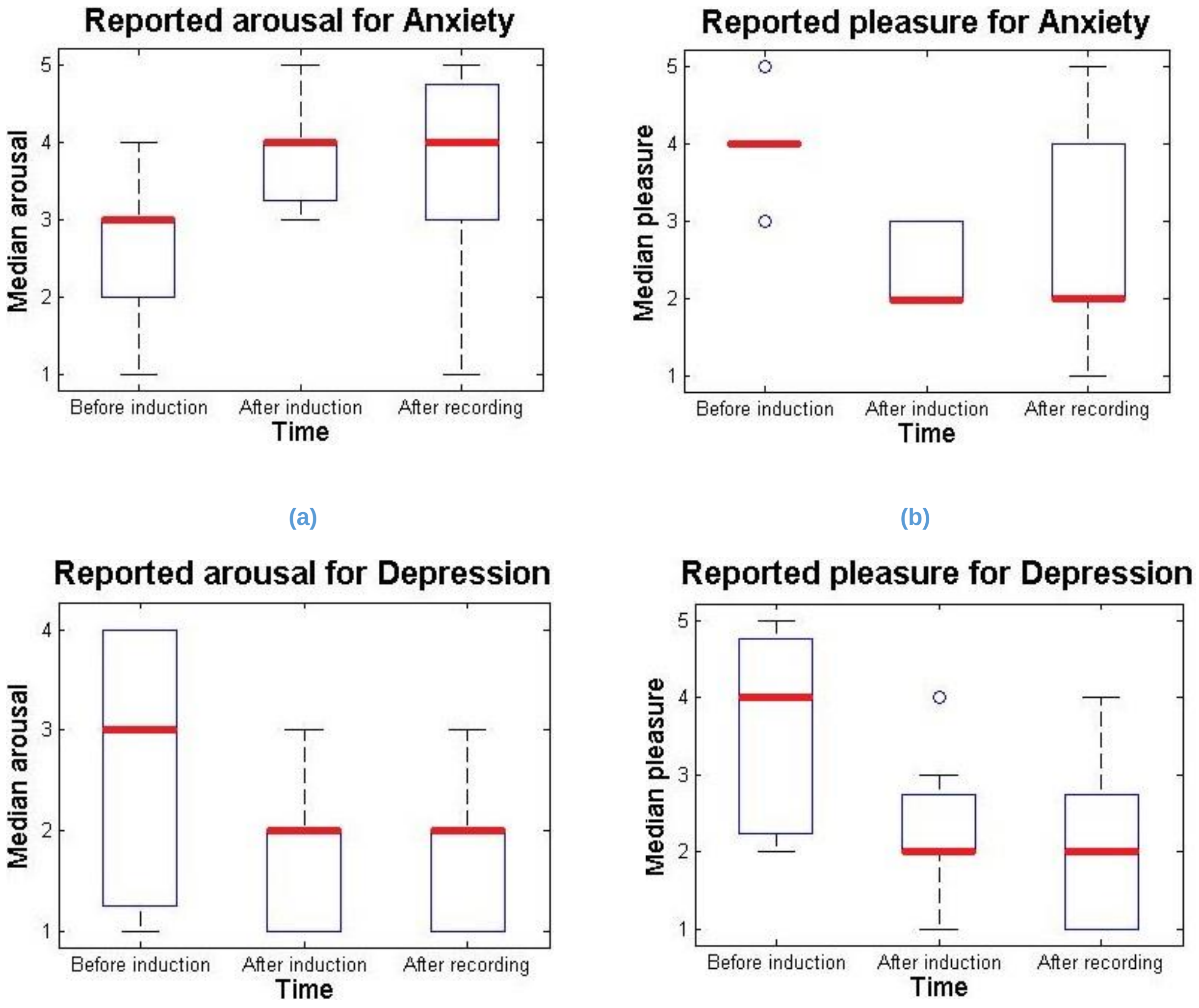
4.4.2 Induction Results

We investigated whether the induction was successful; in other words, if the actors' mood had changed significantly after the induction. For each mood and each reported dimension (pleasure and arousal), we performed a Wilcoxon Signed-rank test using the time of the SAM measurement as the independent variable and the mood rating as the dependent one. We compared, across actors, the SAM scores from time 1, before the induction, with the SAM scores from time 2, after the induction, and the SAM from time 2 with the SAM from time 3, after the recording.

For Anxiety, there was a significant median change from time 1 to time 2 in the levels of pleasure ($Z = -3.493$, $p < .001$) and arousal ($Z = -3.275$, $p = .001$). Pleasure dropped and arousal increased, as intended. See Figure 4-3a and b. For Depression, there was a significant median change from time 1 to time 2 in the levels of pleasure ($Z = -2.274$, $p = .023$) and arousal ($Z = -3.493$, $p < .001$). Both dimensions decreased, as is to be expected for depression. See Figure 4-3c and d.

Amusement was a mood used as a counterexample to Depression and Anxiety, the actual moods that should be recognized by the recognition system. As our focus lay on these moods, no SAMs were recorded for Amusement. In future work, it is interesting to perform these analyses for the induction of amusement as well.

No significant change was found from time 2 to time 3 for any of the moods, for any of the dimensions. This indicates that over the course of recording, the moods the actors were in (i.e., the induced moods) did not significantly change: it would seem that the use of music to induce mood during the recording as well was effective.



(a)

(b)

(c)

(d)

Figure 4-3: Boxplots of the median reported SAM values before mood induction, after mood induction and after the recording, sorted by intended mood and reported dimension.

4.5 Conclusion

This chapter has described the design of a video database using induced and acted moods, and the process of recording the videos. The database has been designed in three steps. We performed interviews at a care home to get information on the situations that give cause for negative moods, and to get physical cues which caretakers interpret as signs of negative moods. These situations led to scenarios for the actors to perform, and the physical cues led to instructions for the actors, which they could use with physical acting techniques to adopt the physique of a person in the given mood. We used these scenarios and physical cues as material for the actors, and induced the mood using music induction. A Wilcoxon test revealed that the actor's experienced mood was significantly different after the induction than before the induction. Specifically, for both sadness and anxiety the pleasure ratings significantly decreased during induction, as expected; arousal ratings decreased after sad mood induction, whereas, again as expected, they increased after anxious mood induction. Furthermore, the analysis revealed that the induced mood stayed constant throughout the recording phase, suggesting that the actors, although not explicitly instructed to portray a specific mood, may have conveyed it appropriately and in a naturalistic way, and that the induction was successful.

The product of this step was the raw video data for the database, with facial data, data on the whole body, and skeleton data from the Kinect.

Chapter 5: Crowdsourced Database Annotation

The training of the affect recognition system this database is intended for requires, in a supervised learning setup, not only recordings of people in different moods are necessary, but we also need to provide the system with an indication of the mood that is portrayed in each video. We have focussed our research on the physical display of mood. While we have verified a consistency between the mood the actors felt and the mood we intended to induce, we have not yet verified whether the physical *display* of mood is perceived by humans as intended: it is important to recall, once again, that actors were not explicitly asked to portray specific moods: they were only provided with scenarios, and underwent mood induction. To be able to determine, for each video, the mood portrayed in it, we need thus to resort to external, human annotators. This chapter describes the process of having the videos annotated by external observers to acquire this ground truth.

The typical procedure to annotate affective videos is to rely on a relatively small number of experts (Abadi, et al. 2014, Snel, et al. 2012). For example, the HUMAINE (Douglas-Cowie, et al. 2007) and SEMAINE (McKeown, et al. 2012) databases have at most six expert annotators per video, and Creative-IT (A. e. Metallinou 2010) and IEMOCAP (Busso, et al. 2008) have no more than three. However, in our case we had sixteen hours of video material from the two cameras. The annotation of such a large amount of video material by experts would be unfeasible because of (1) the limited time resources of the experts and (2) the risk of fatigue and thereby unreliability of the annotations. As an alternative, we looked at crowdsourcing, which involves decomposing the large annotation task into microtasks, to be outsourced to a large amount of naïve annotators (the crowd). We expected that relying on the crowd wisdom would be effective, as crowdsourcing typically gives low costs and high completion speed (Hsueh, Melville and Sindhwani 2009). Nevertheless, the questions (1) whether a large number of naïve annotators could substitute a handful of experts and (2) whether the highly uncontrolled environment in which crowdsourcing tasks are executed would pose a limit to the reliability of the annotations, were major concerns and therefore core objectives of our investigations.

The remainder of this chapter is organized as follows. We will first discuss the goal and scope of the annotation task. We will then describe how we designed the task, followed by the mechanisms we employed to control the reliability of the results. Next, we summarize the annotation protocol from the annotator’s viewpoint. Finally, we give the technical details of the implementation of our task.

5.1 Overview of the crowdsourcing annotation campaign

In this experiment, we wished to investigate whether crowdsourcing is suitable for annotating an affective database. In this section, we describe which measurements we were seeking, for which videos, by whom.

5.1.1 Crowdsourcing Platform

Introduced in Chapter 2, a crowdsourcing platform connects employers to workers. The crowdsourcing platforms we saw most commonly used for affective annotation are Mechanical Turk, Microworkers and CrowdFlower. All three are used in both commercial and research contexts. Mechanical Turk may only be used in the USA, which made it impossible for us to use. CrowdFlower offers built-in tools to help a crowdsourcing employer in the collection and checking of the acquired data, but in turn requires that the application containing the task is written in CrowdFlower's own markup language (CrowdFlower 2015). When using Microworkers, on the other hand, the connection between platform and application is much looser: the campaign on the platform must simply contain a link to the site where the workers can find the application. As will be discussed in Section 5.4 , we were basing our application on an existing site, written in HTML. The most efficient choice, therefore, was to use Microworkers so we could expand Soleymani's existing application without having to translate or change it to adapt to a specific crowdsourcing platform.

5.1.2 Measurements

The goal was to annotate the emotions and mood as the perceived affective states of the person acting in the video. We described these affective states using two dimensions: arousal and pleasure (Russell 2003). See Section 2.1 for more details. We required two kinds of annotations:

1. *Continuous annotation of emotion*, expressed as a trace of perceived pleasure or arousal of the person in the video, throughout the video. This annotation was performed while watching the video, using a trace tool such as described in Section 2.1.2.2 .
2. *Global annotation of Mood*, expressed, after watching the whole video, as a single value for pleasure or arousal summarizing the overall mood of the person in the video. This value was measured after watching the video, using the SAM, to allow comparison with the SAMs filled out by the actors in the video.

5.1.3 Task setup and video material

From our recording experiment, we had videos from the camera zoomed in on the face of the actor, videos from the camera focussed on the whole body, and recording data from the Kinect. We expected that, on the smaller screen of our website, mood would be easier to recognize from the face videos. Therefore, we started with the annotation of the face videos. Browsing the existing jobs on Microworkers, we concluded that most tasks take no more than five minutes to finish, which coincides with the literature (Hirth, Hoßfeld and Tran-Gia 2011). We decided to use five minutes as the maximum length of one task, annotating one video continuously and then annotating the mood. We estimated the mood annotation would take one minute, leaving four minutes for the video watching (including continuous annotation). This meant we had to cut the videos that were longer than four minutes. Additionally, since most actors stood up multiple times in the videos, moving outside the capture range of the face camera, those parts of the video had to be cut out. Videos were edited using Adobe Premiere Pro. We cut out the parts where the actor was not on-screen, and we split up videos that were longer than four minutes.

The resulting material consisted of 180 clips showing the face of the actors, each intending to depict one of the following three moods: anxiety, depression and amusement. As it is good practice to have the payment for a task depend upon the amount of necessary effort (Gadiraju, Kawase and Dietze 2014), we divided the videos based on length. The videos were split into two groups: 92 'long' videos longer than two minutes, and 88 'short' videos shorter than two minutes.

5.1.4 Participants

To investigate the reliability of the acquired annotations, we needed an appropriate amount of annotators to achieve a reliable measure of agreement: a measure that would indicate an acceptable level of agreement in the cases where annotators did indeed agree. Cowie and

McKeown (Cowie and McKeown 2010) investigated how many raters are needed to get acceptable levels of agreement as calculated by Cronbach's alpha. They reported the relationship between average correlation and alpha as a function of the number of raters, and found that with six raters, a correlation of only 0.30 is needed to reach the standard acceptable level of alpha, 0.7. Therefore, we wanted to get *at least* six annotators per video. Since in a crowdsourcing environment we would likely have to discard a number of unreliable results, we decided to aim for ten annotators per video, amounting to 1800 annotators in total (assuming one annotator per video).

5.1.5 Campaigns

We divided the annotators into four groups, grouped by two factors: dimension and length of video. The dimensions to be annotated were pleasure and arousal. The four groups are listed in Table 5-1. For each group, we submitted a different Microworkers campaign. We decided on the division between dimensions to prevent having participants watching the same video twice. We decided to group by length of video to allow different payments according to length of the video, and therefore duration of the task. Payment was decided based on observed payments from existing jobs and payments from comparable jobs in the literature. We observed from existing jobs on Microworkers that the average hourly pay was approximately $2.00 per hour. Most of the tasks were simple tasks such as signing up for a website. Soleymani et al. (Soleymani, Caro, et al. 2013) paid $3.12 USD per hour in a similar experiment for annotating music. Hsueh et al. (Hsueh, Melville and Sindhwani 2009) paid four cents for an annotation task during 40 seconds, which is $3.60 per hour. We decided to pay $0.30 for annotating one of the longer videos (> 2 minutes), and for a shorter video we paid $0.20. This amount was slightly higher than the simpler tasks on Microworkers, making it stand out from those tasks.

1. Arousal videos shorter than two minutes	2. Pleasure videos shorter than two minutes
3. Arousal videos longer than two minutes	4. Pleasure videos longer than two minutes

Table 5-1: The four groups of campaigns. The participants were grouped by dimension and length of videos.

The database included a large amount of videos, which meant we would get a large amount of results back at once, all of which had to be checked and paid manually. To ensure we could react to workers in time, we had the database annotated in five batches, three long and two short. Each batch was a subset of the original database, containing the maximum possible diversity in moods (depression, anxiety and amusement) and actors. Once one batch was completely annotated along both dimensions, we put the next batch on Microworkers.

5.2 Reliability Control

As discussed in Section 2.3.3 , a main concern in using crowdsourcing for scientific research and specifically for the annotation of media material, is the reliability of workers in task execution. Lack of comprehension of the task, poor instruction or task design, lack of motivation of the workers to perform the annotations properly can hamper the reliability of the data collected, and as a consequence, endanger the reliability of the affect recognition systems trained on those data.

We evaluated different strategies to control and verify the reliability of the workers and their annotations, and specifically we implemented verification and consistency tests, content questions, and application-layer monitoring. We made the controls in such a way that annotation was allowed to continue only while the participant was actually operating the controls. This feature will be explained further in Section 5.4. We used a qualification test before the application started, using gold standard data from a small pilot. We implemented a number of reliability control mechanisms for data that would help us filter out the unreliable workers post-hoc: we used consistency tests, content questions and application-layer monitoring of the task time. As it is advisable not to include too many mechanisms to prevent the task from becoming tedious (Hoßfeld, Seufert, et al. 2011), we did not include a common knowledge question. In this section, we will describe these mechanisms in detail. The analysis we performed based on these mechanisms will be explained in Chapter 6.

5.2.1 Qualification test

Before allowing the participant to perform the annotation, we wanted to test whether he/she understood the task and had the empathic abilities required for an affect recognition task. This is a complex task, however, as emotional perception is a very subjective matter (Y. Baveye, et al. 2013). We showed the participant two videos already annotated from a small scale pilot study, and after each video we asked him/her to rate it using a SAM. Each video was seven seconds long. For the pleasure task, we showed a video of an amused person (positive pleasure) and a video of a depressed person (negative pleasure). For arousal, we showed a video with a depressed person (low arousal), followed by a video with an anxious person (high arousal). We checked whether the participant’s answer was in the same range as the data we had received from the pilot study. If so, the participant was allowed to proceed to the annotation task. If not, they were blocked from the experiment. Blocked workers were not paid.

5.2.2 Content Questions

This step aimed at verifying that the annotators paid full attention to the video. At the end of each annotation, the participant was asked four questions about the actor they had seen in the video. We wished to devise a set of questions that were centred on actions that appeared in some, but not all of the videos, and that should be clearly seen by an attentive viewer. This proved difficult, as in most videos, movements were subtle, and most actions were improvised by the actors. We therefore based the questions on the actions dictated by the scenarios. A disadvantage in this approach is that, as all questions were the same for all videos, a participant rating more than one video might learn to pay attention to the details requested in the questions and to nothing else. We relied though on the hypothesis that the other filtering mechanisms would have identified most of these less-than-reliable workers.

The questions asked to each participant at the end of the video were:

1. What was the person’s gender? (M/F)
2. Was the person eating? (Y/N)
3. Was the person having a telephone call? (Y/N)
4. Was the person reading? (Y/N)

The participant could select the right answer via a radio button.

5.2.3 Task duration

We used task duration as a further indicator of reliability in task execution. Participants that took long to annotate their video might have paused halfway through before continuing. To filter out these people, we kept track of the time it took the participant to annotate a whole video. We could later compare this time with the video length: we removed annotations where the completion time was more than three standard deviations above the mean completion time for that video.

5.2.4 Consistency

As key control step, we implemented a mechanism to verify that workers would be faithful in their evaluations of mood. To do so, we had workers annotating mood using, sequentially, three different tools: a SAM first, then an AffectButton, and finally a descriptive categorization of the mood. For pleasure, the possible categories were: positive, neutral and negative. For arousal the categories were: high, medium and low. This setup would allow us to filter out any participants whose answers differentiated too much from each other on the three different tools.

5.3 Annotation Protocol

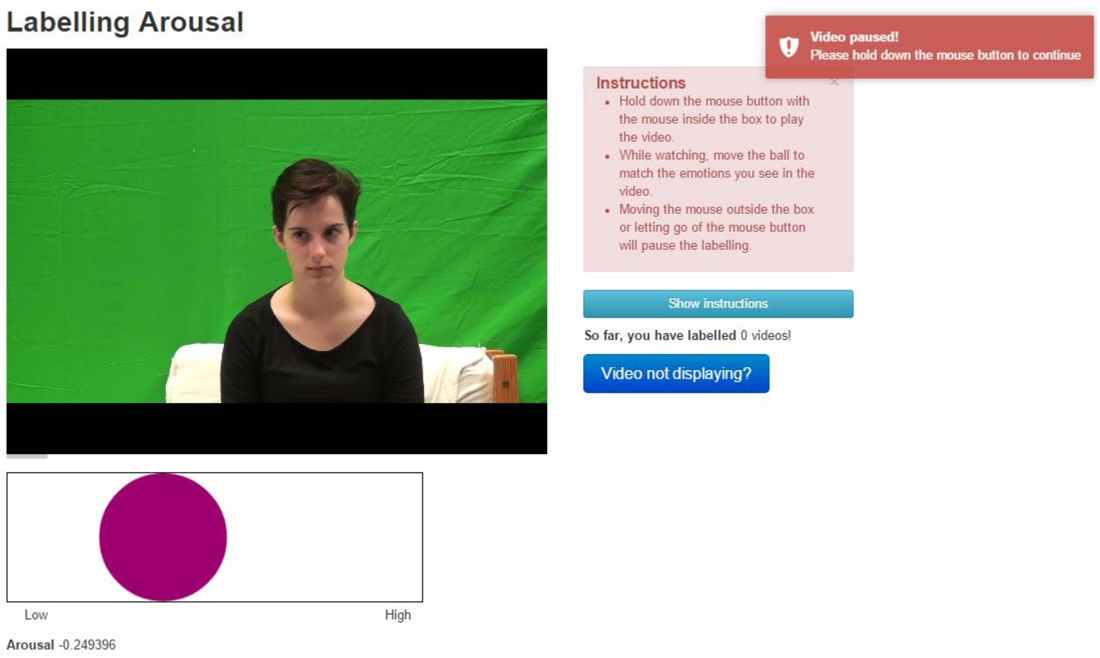
The participant could perform the annotation tasks on his/her own computer, using Firefox 3 or later or Chrome 7 or later. Via Microworkers, the participant followed a link to a website containing our application (the implementation of which will be described in Section 5.4). Upon opening this site, the participant was greeted by a welcoming screen describing the task. This screen also explained the meaning of the dimension (pleasure or arousal) the participant was annotating, and the use of the SAM. The participant was asked to fill out his/her gender and ID from Microworkers. The participant was asked with an informed consent form to agree that they were jointing the experiment voluntarily, that the data would be treated anonymously, that they could stop at any time, that they would participate in the experiment seriously and that they would be paid for each fully annotated video.

The participant was then asked to perform the qualification test described in Section 5.2.1 . If the participant passed the test, he/she was considered qualified to proceed with the affective annotating, otherwise he/she was kindly informed that he/she was not in the position to complete the annotation task. Disqualified subjects were not paid.

If qualified, the participant viewed a short tutorial that explained the functionality of the annotation screen (see Figure 5-1) and specifically the slider. After the tutorial, the actual annotation started. The application loaded a video. The participant clicked on the slider to start the video and the annotation. He/she held down the mouse button for the entire length of the video, and moved the slider along the axis of the dimension he/she was annotating. When the video was completed, the application moved to the next screen, asking him/her to evaluate the mood of the person in the video by filling out a SAM. See Figure 5-2. Since our scenarios contained mood shifts, it was possible that the video contained two moods. Therefore, the participant was asked whether he/she had seen one or two moods. In the case of two moods, the participant was asked to fill out a SAM for each mood.

As a part of the consistency check, the participant then filled out a mood evaluation using the AffectButton and afterwards using a categorical description. Using the AffectButton, the participant was asked to experiment with the button and explore its emotional ranges, and then adjust the face to match the mood they had perceived and press the button. For the AffectButton and the categorical rating, if the participant had perceived the person in the video to be in two different moods throughout the video, we specifically asked to evaluate only the last recognized mood. The participant was also asked the content questions, and was then allowed to submit any comments. Having completed the annotation for this video, the participant was asked if he/she wanted to annotate another. If so, the site loaded the next video and the process was repeated. The experiment lasted five minutes per video on average.

At the end of the experiment, after the participant had annotated up to ten videos, the participant was allowed to leave a general comment and given a verification code which could be submitted to Microworkers to claim payment. The verification code consisted of four parts: the participant's hashed Microworkers ID, the amount of videos annotated, the title of the last video and the annotation time. These four parts could later be checked to see if they matched up with the database. If so, the participant was paid $0.30 for videos longer than two minutes and $0.20 for shorter videos.



Labelling Arousal

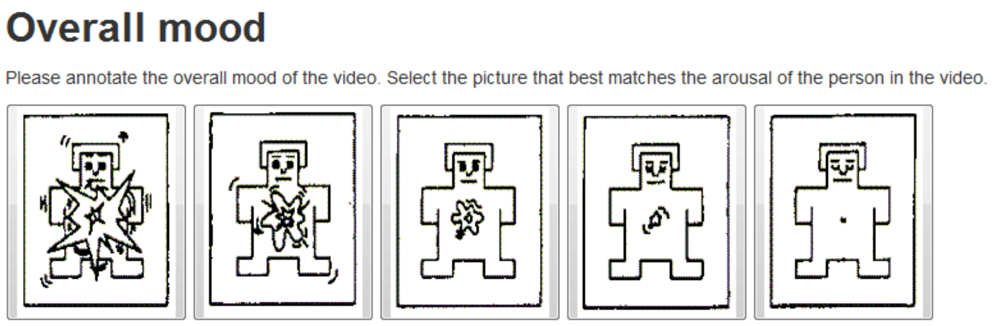
Video paused!
Please hold down the mouse button to continue

Instructions
- Hold down the mouse button with the mouse inside the box to play the video.
- While watching, move the ball to match the emotions you see in the video.
- Moving the mouse outside the box or letting go of the mouse button will pause the labelling.

So far, you have labelled 0 videos!

Arousal -0.249396

Figure 5-1: Screenshot of the site for continuous annotation. The participant views the video in the top-left corner while annotating it using the slider in the bottom-left corner. The video does not play unless the mouse is clicked and is inside the rectangle box.



Overall mood

Please annotate the overall mood of the video. Select the picture that best matches the arousal of the person in the video.

Figure 5-2: Screenshot of the SAM popup screen for arousal. The participant selects the manikin that best matches the affective state they perceived in the video.

5.4 Annotation application

The application used for this annotation is a modified version of Soleymani's online application (Soleymani, Caro, et al. 2013), inspired by tools such as GTrace (Cowie and Sawey 2011). The application was created in HTML and JavaScript. We edited the application to support videos rather than music, added our reliability control mechanisms, edited the GUI according to the experimental procedure we had defined and added created a MySQL database for storing the results. This section contains the technical details behind these modifications.

5.4.1 Tools

The original tool of Soleymani (Soleymani, Caro, et al. 2013) uses jPlayer (Ltd 2015) to play music. This tool can also be used to play videos. We set up the player to play videos and adjusted the layout accordingly. The application requests the list of video names from the database belonging to the videos that have not yet been fully annotated, shuffles this list and sets it as the player's playlist. Holding down the mouse button while the cursor is in the slider area triggers the annotation and causes the video to play. Letting go will pause the video as well as annotation. For the continuous annotation, we used the slider already implemented in the application. The slider is a rectangular canvas with a circle drawn inside, as seen in Figure 5-1. While annotating, the video keeps track of the video time and the position of the slider. These values are appended to an array every 250ms during playback.

The *content questions* can be answered using a set of radiobuttons. The answers are collected and sent to the databases upon confirmation.

Time is collected using the JavaScript Date object. This data is collected once the participant loads the page, when the participant starts annotating a video, when the video is ended and when the participant submits the annotations.

For the mood annotation, we used three different tools to check for *consistency*. Primarily we used the SAM, the same tool we used with the actors. This would allow comparison with the mood reported by the actors after the recording of the video. We added the SAM in the form of radio buttons with the five SAM icons as images.

We implemented the AffectButton by including the libraries provided on the AffectButton's site. (Broekens, Joost Broekens 2015) After clicking on the AffectButton, the smiley freezes in its place and the participant is asked if the current facial expression is the mood he or she had in mind. If so, the values are saved. If not, the AffectButton becomes active again and the participant can select a new face.

The Categorical annotation is asked along with the content question, as a separate screen would make the task unnecessarily longer. It uses the same radio buttons.

The start-up *test* uses the jPlayer tool together with the SAMs as used in the mood annotation.

5.4.2 MySQL Database

We set up a MySQL database server, and extended the tool with PHP scripts which saved the data from the experiment in the MySQL tables. Figure 5-3 shows the database structure. The database contains six tables:

- In the *sam* table, each row contains data of a single mood annotation. Each row contains a unique identifier, the video which was annotated, the rater's hashed id, the dimension (arousal or pleasure), the modality (face or body), any comments left by the annotator, the sam which is mapped to a number from 1 to 5 (1 being the manikin with the lowest value, 5 the highest), the values for the Affect Button, the answer to the questions and four timestamps.
- The *slider* table contains all the continuous annotation data. Each row contains the position of the slider at a given time. Aside from the value, the in-video time is stored, as well as the amount of time that has passed since the video started and the matching identifier from the *sam* table.
- The *raters* table contains the data on the participants. A hashed version of their Microworkers ID and gender is stored, as well as the number of annotations, and the score achieved in the start-up test. The test consisted of two videos, and this score

indicated which videos had been given the expected evaluation. If this number was lower than two, the participant was not allowed to participate.

- The *videos* table contains the data on the videos. The number of annotations per video is stored here in the 'annotations' column, and the number of participants that are currently annotating this video in the 'reserved' column. When the participant starts annotating a video, the system increments the 'reserved' counter for that video. When the participant submits his/her annotation, the system decrements the number and increments the 'annotations' counter. This way, the system 'reserves' a video for a participant. To prevent an excessive amount of participants annotating one video, the system stops loading a particular video when the sum of the two counters reaches ten. Additionally, this table contains the correct answer to the content questions for each video: whether the actor is seen eating, reading, making a call and the actor's gender.
- The *comments* table contains any additional comments at the end of the experiment. It stores the hashed Microworkers ID of the annotator who left the comment, as well as the comment itself.
- The *log* table is used to keep track of any errors that may have occurred, as well as all SQL queries that have been sent. Its data is used as a backup as well as a debugging log. It contains a timestamp, the participant's hashed Microworkers ID, the identifier from the sam table (only if it has been saved in the sam table already), and whether the submission comes from an event that occurred in the JavaScript code, the PHP code or the SQL query.

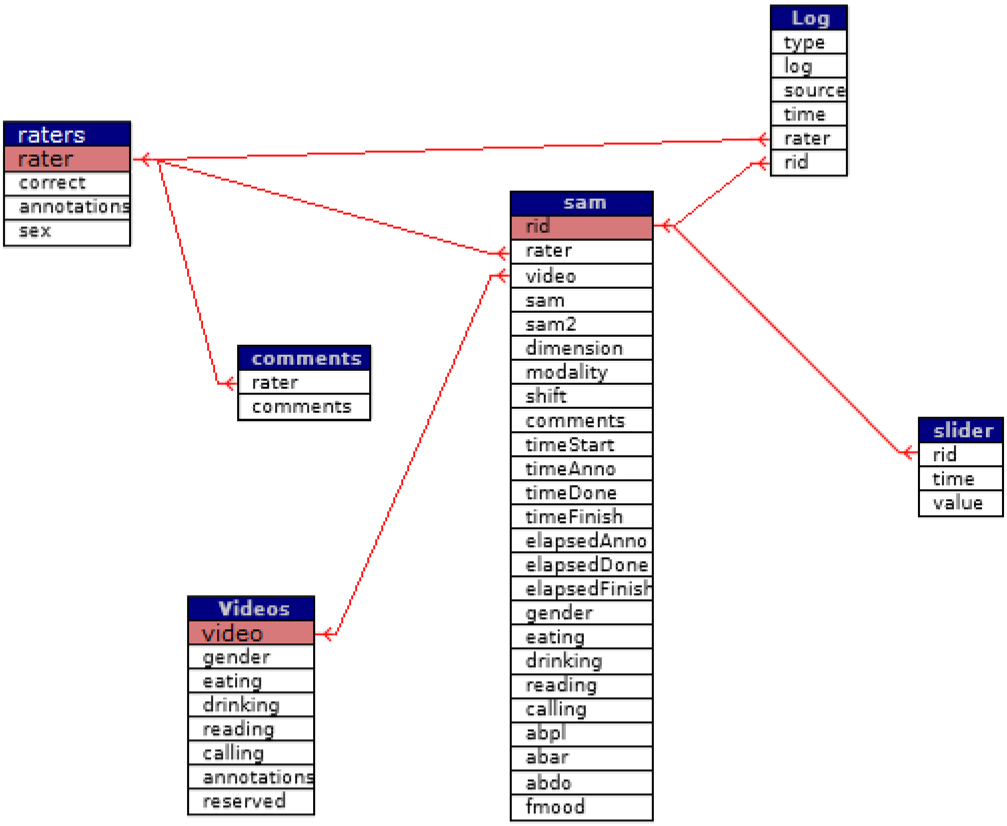


Figure 5-3: Database relational diagram. Primary keys are displayed in red.

5.5 Results

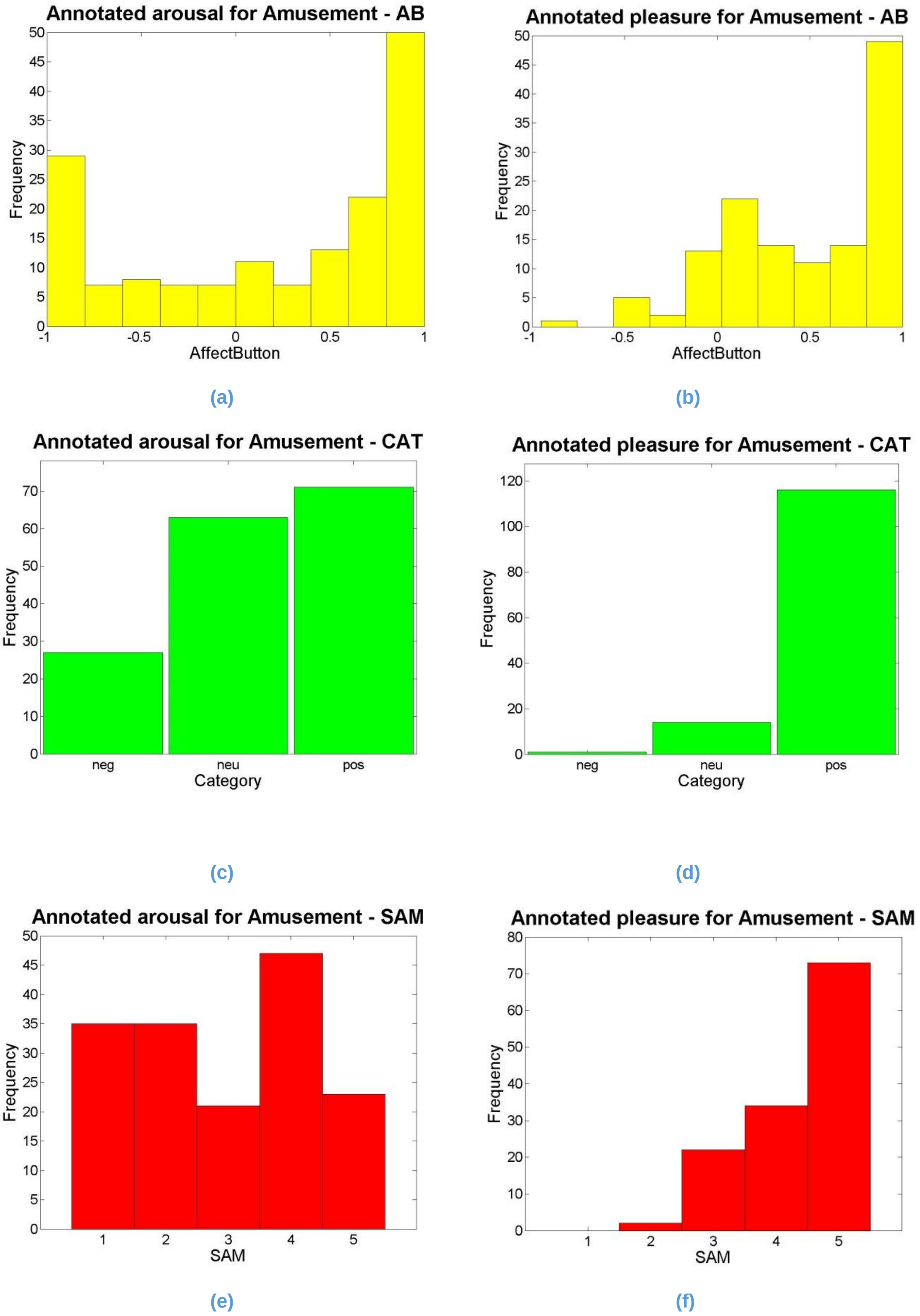
We collected 2150 mood pleasure annotations (with 613 different annotators) and 2424 (with 451 different annotators) for mood arousal. The task took 5.6 minutes on average for pleasure and 5.2 minutes for arousal. We spent $1500 on the campaign. To give an indication of the results, we display the histograms belonging to each tool and each intended mood, followed by the histograms for the continuous emotion annotations sorted by each intended mood, in the figures below.

5.5.1 Mood annotation

The histograms for the static mood annotations are displayed in Figure 5-4 to Figure 5-6. As can be seen, the annotation distributions have their peaks at the expected places: high pleasure for amusement (see Figure 5-4.b – Affect Button –, 5-4.d – categorical – and 5-4.f - SAM), low pleasure for anxiety (Figure 5-5.b, d and f) and for depression (Figure 5-6). The arousal for anxiety has mostly high values for the SAM (Figure 5-5.e) and the categorical description (Figure 5-7.c), but less so for the AffectButton (Figure 5-5.a). This could be explained by the fact that for the Affect Button, the arousal values are derived rather than directly influenced by the participant when moving the button (Broekens and Brinkman 2013). More specifically, when the mouse pointer is held around the middle of the AffectButton, the facial expressions are more subtle than at the outer areas of the button. These expressions give low values of arousal. The annotators might have selected these more subtle faces to match the subtle displays they saw in the actors (given that we focussed our work with the actors on bodily expression of mood rather than the facial expression), explaining why we see such a high peak at the lowest ranges of arousal for anxiety in Figure 5-5a.

5.5.2 Emotion Annotation

The histograms for the continuous emotion annotations are displayed in Figure 5-7. The majority of the emotions for amusement have generally been given high ratings for both arousal (Figure 5-7.a) and pleasure (Figure 5-7.b), as expected. The arousal for anxiety (Figure 5-7.c), however, is peaked around 0, which is lower than expected; the shape of the histogram appears to follow a normal distribution. We see a similar occurrence with the pleasure for anxiety (Figure 5-7.d) and the pleasure for depression (Figure 5-7.c) where the means are close to 0 and higher than expected. We see a generally low arousal for depression (in Figure 5-7.e), which is expected.



(a)

(b)

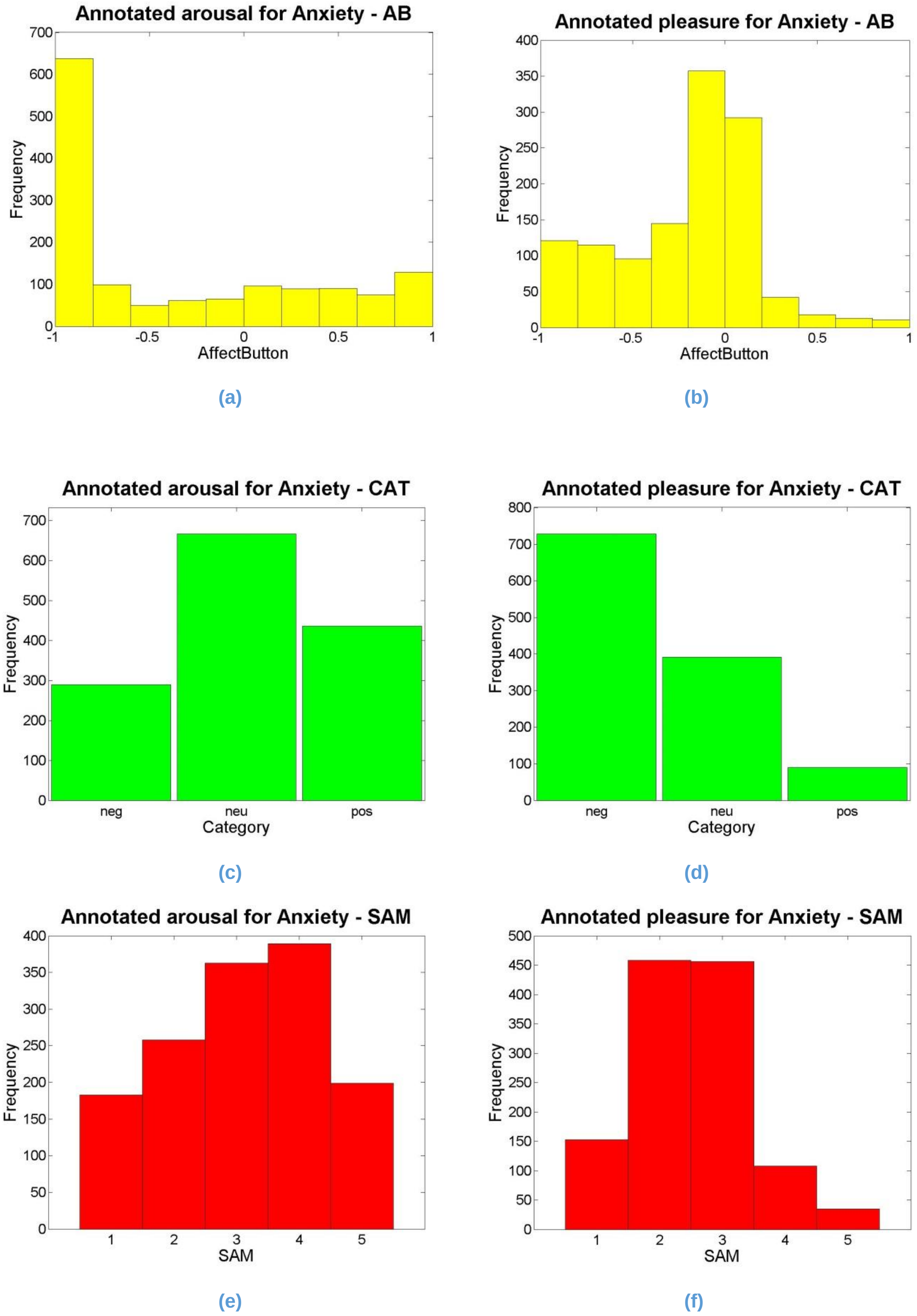
(c)

(d)

(e)

(f)

Figure 5-4: Histograms of the observed values for the videos intended to show *Amusement*. A histogram is displayed for pleasure (right) and arousal (left), for each static mood annotation tool: Arousal (a) and pleasure (b) values from the AffectButton (AB) in yellow, arousal (c) and pleasure (d) values from the categorical scale (CAT) in green, arousal (e) and pleasure (f) values from the SAM in red.

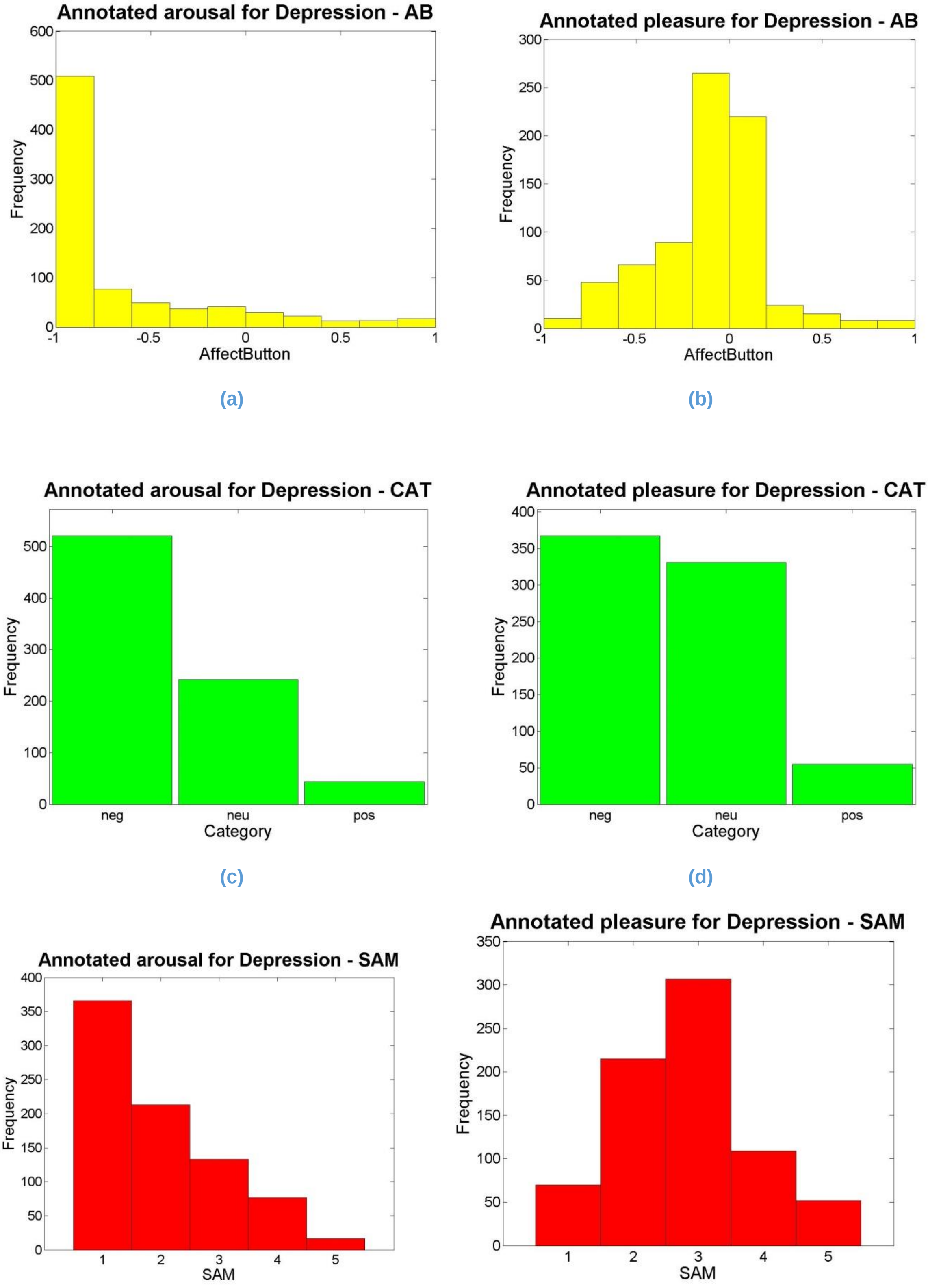


(a) (b)

(c) (d)

(e) (f)

Figure 5-5: Histograms of the observed values for the videos intended to show *Anxiety*. A histogram is displayed for pleasure (right) and arousal (left), for each static mood annotation tool: Arousal (a) and pleasure (b) values from the AffectButton (AB) in yellow, arousal (c) and pleasure (d) values from the categorical scale (CAT) in green, arousal (e) and pleasure (f) values from the SAM in red.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5-6: Histograms of the observed values for the videos intended to show *Depression*. A histogram is displayed for pleasure (right) and arousal (left), for each static mood annotation tool: Arousal (a) and pleasure (b) values from the AffectButton (AB) in yellow, arousal (c) and pleasure (d) values from the categorical scale (CAT) in green, arousal (e) and pleasure (f) values from the SAM in red.

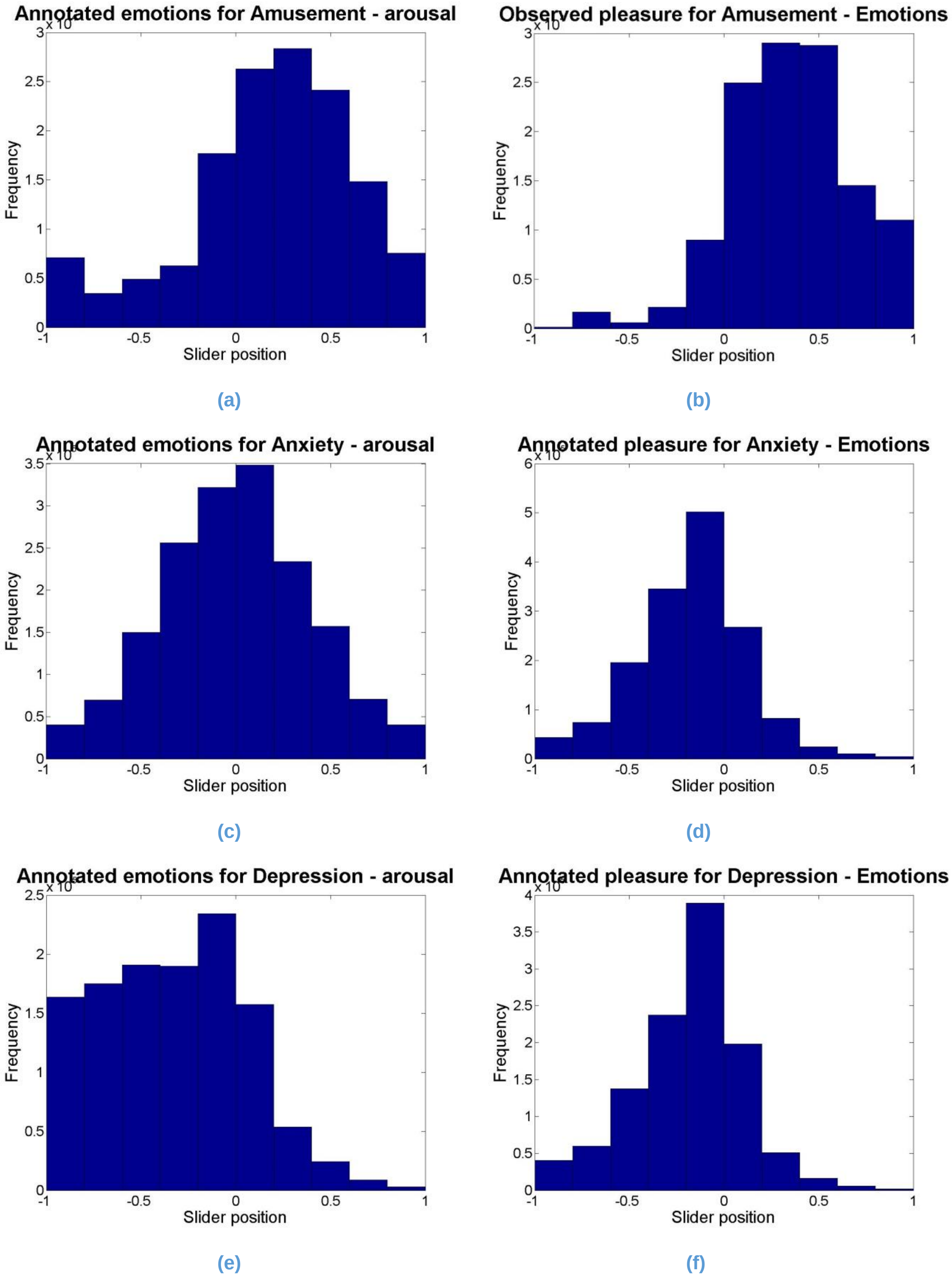


Figure 5-7: Histograms of the observed values of *emotion* at all times for the slider sorted per intended mood. A histogram is displayed for pleasure (right) and arousal (left), for each intended mood: Arousal (a) and pleasure (b) values for Amusement, arousal (c) and pleasure (d) values for Anxiety, arousal (e) and pleasure (f) values for Depression.

5.6 Conclusion

We used crowdsourcing to get two types of annotation: mood annotation, which was annotated using the SAM after watching the whole video, and continuous annotation, which was annotated dynamically using a slider while watching the video. To reach this goal, we developed an application based on Soleymani's tool (Soleymani, Caro, et al. 2013) for annotating music.

After running this experiment on Microworkers, we had annotation data for all the videos in the database. As a crowdsourcing environment is prone to unreliable worker behaviour, we built in several ways to check the reliability of the data we received from the experiment. The next step would be to use these checks to filter out the annotations that were deemed unreliable.

Chapter 6: Analysis of the annotations

We have collected insights on non-verbal mood expressions of elderly care-takers, recorded actors incorporating those insights into their acting, and had the resulting material annotated via crowdsourcing. In this chapter, we examine the results from these endeavours in order to answer the research questions we started with.

Firstly, we consider the three intended moods: depression, anxiety and amusement. We wish to verify whether these three moods have indeed been recognized by the annotators. Secondly, we want to examine how reliable the annotations were, and to what extent we are able to control the reliability of the annotations. We have described the different reliability control mechanisms we implemented in the crowdsourcing campaign in Chapter 5. In this chapter, we use the data from these mechanisms to filter the annotations, removing any annotation that does not meet the criteria of the mechanisms.

This chapter consists of four sections. In Section 6.1 we explain how the reliability control mechanisms were used to filter the data. In Section 6.2 we explain which analyses we performed on the data, comparing the annotated mood with the intended mood and calculating the inter-rater agreement to examine how reliable our annotations are. We present the results for each filtering stage, and examine the filtering process: we investigate how each choice of criteria affects the number of discarded annotations and how the quality of the remaining annotations is affected. Having considered each option, we choose a set of filtering criteria in Section 6.3 where we then examine the results of the analysis belonging to that set of criteria.

6.1 Filtering

As reported in Section 5.2 , we implemented a number of mechanisms to control, pre-test and post-hoc, the reliability of the task execution and thereby of the annotations. This section presents the results of the process of removing annotations that were deemed unreliable according to these checks. We define this process as filtering the annotation data.

Each reliability control mechanism results in a filtering stage, in which we discard a number of annotations if they do not meet a criterion imposed by that reliability control mechanism. At every next filtering stage, fewer data passes through, the quantity of which can be adjusted by selecting a “strict” or more “lenient” criterion. To allow the comparison of the effects of filtering step-by-step, we display the results of each stage of filtering in a decision tree. For each stage, we show reliability and statistical measurements for the annotations remaining after applying that filter and the previous stages of filtering.

We applied four filtering stages: one based on the analysis of the (continuous) emotion annotations, one based the content questions, one based on task duration and finally one based on the consistency of the mood annotation. These stages are summarized in Section 5.2 .

6.1.1 Emotion annotations

This stage aims to filter out annotations where the participant did not annotate actively. We analysed all continuous annotations, and checked how often the position of the slider changed after the first five seconds of annotation (we used a buffer of five seconds to account for the need to get acquainted to the slider usage). If the participant did not move the slider at all, we assumed that the participant had not paid attention to the affective changes in the movie or did not bother to annotate them. Therefore, we only accepted annotations where the slider was moved more than once after the initial five seconds.

6.1.2 Content questions

This stage concerns the questions asked at the end of the annotation to check if the participant had paid proper attention to the video. There were four questions. However, from comments left by the participants, we discovered that the last question (“Was the person reading?”) was ambiguous. The videos intended to portray amusement, for example, showed actors watching a movie on an iPad. This could be mistaken for a person reading an article on the iPad. In other videos, an actor took a quick glance at a paper. Whether this counts as reading is disputable. For these reasons, we disregarded the fourth question.

We used the remaining three questions as a criterion for filtering the annotations. In this case we differentiated the strictness of our filtering in two options. With *strict filtering*, we accepted only annotations where all three questions were answered correctly. With *lenient filtering*, we accepted only annotations where at least two questions were answered correctly.

6.1.3 Annotation time

We examined the amount of time each annotation took, from the start of the video to the end of the video, after subtracting the length of the video, resulting in the *relative completion time*. We then removed annotations that took too long, calculated as follows:

$$discard\ if\ relative\ completion\ time > (3 * \sigma + \ \mu)$$

where σ and μ are the standard deviation and mean, respectively, of the relative completion times for all annotations for that dimension (pleasure or arousal). This is a rather lenient criterion, as often in literature outliers are detected as those deviating from the mean for more than 1.5 times the standard deviation. However, because it is possible that our participants took longer to finish the video because of a bad internet connection, we decided to only apply a more lenient filtering criterion with respect to time, and accept all annotations where the annotation time was smaller than three standard deviations above the mean completion time.

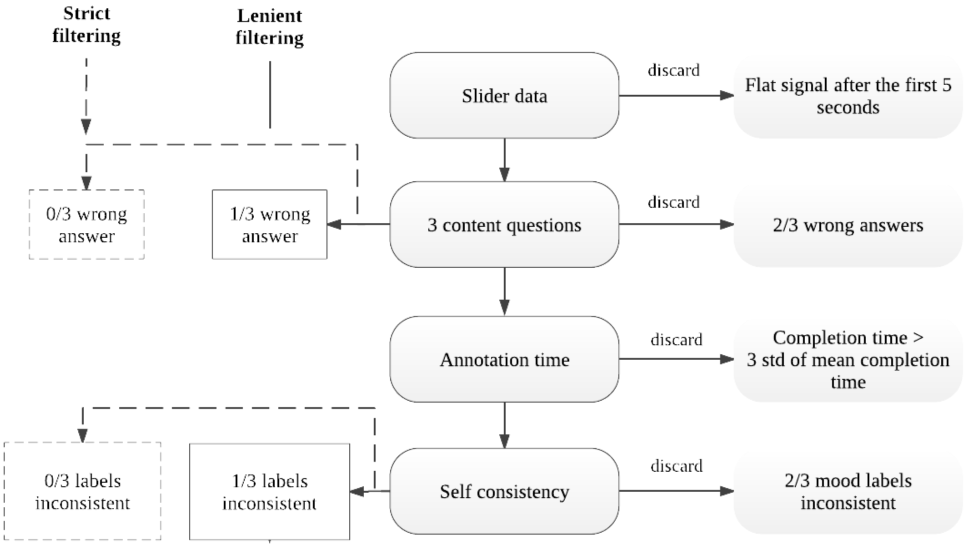


Figure 6-1: Graphic representation of the filtering process. If a criterion is not met, the full annotation is discarded.

6.1.4 Consistency

To test the consistency of the participants, we asked them to rate the mood using three different tools: the SAM, the AffectButton and a three-point categorical scale. The SAM uses a five-point scale and the AffectButton returns a continuous value between -1 and 1. To be able to compare these values, it was necessary to map the values of one scale to another. However, simply applying a linear transformation to these values (e.g. normalizing the categorical values in the rage [1, 5] of the SAM) was not a reliable solution, as we cannot assume the values from the SAM and from the categorical scale to be interval values. In addition, differences in the ratings provided by each tool could be caused by different interpretations of what the scales of each tool represent: we cannot be sure that the highest value of the SAM scale corresponds to the highest value of the AffectButton. Therefore, we used a data-driven approach.

Since the categorical scale is the scale with the least amount of points, we decided to map the SAM and AffectButton values into the categorical scale. We used a *tree classifier* for this, a hierarchical model in the form of a tree which predicts the class of input data according to a discrete function of the input values. Input is classified by navigating through the tree from the root to one of the leaves which each represent the most likely class (Rokach and Maimon 2014). For each dimension, we trained such a tree classifier mapping the data from the SAM to the categorical data, and a classifier mapping the data from the AffectButton to the categorical data. The resulting trees are displayed in Figure 6-2 for Pleasure and Figure 6-3 for Arousal.

The SAM ratings are denoted as an integer from 1 (lowest rating) to 5 (highest rating). As we see, for the SAM pleasure ratings (Figure 6-2a), a value of 3 is mapped to ‘neutral’, below 3 is mapped to ‘negative’ and above 3 is mapped to ‘positive’. Similarly, for the SAM arousal ratings (Figure 6-3a), a value of 3 is mapped to ‘neutral’, below 3 is mapped to ‘low’ and above 3 is mapped to ‘high’. In other words, the middle rating of the five –point SAM scale is considered neutral by the mapping, as one would expect.

For the AffectButton, pleasure ratings (Figure 6-2b) with values between -.0376 and .2006 are mapped to ‘neutral’, lower values are mapped to ‘negative’ and higher values are mapped to ‘positive’. For the AffectButton arousal ratings (Figure 6-3b), values between -.0116 and .4071 are mapped to ‘neutral’, lower values are mapped to ‘low’ and higher values are mapped to ‘high’. It is not surprising that the pleasure values from the AffectButton that are mapped to ‘neutral’ fall in a much smaller range than those mapped to the other two classes. Recall from the results shown in Section 5.5 : the large majority of the values submitted via the AffectButton lies between -.2 and .2. It is understandable, then, that the boundaries between the three classes lie closer to 0 than where they would be if we drew the ranges up in even sizes. For the arousal values from the AffectButton, the majority of the ratings were close to -1, which may explain why the range of values mapped to ‘low’ (below -.0116) is much larger than the other classes.

For each tree, the classification error was calculated as the *resubstitution loss*. This measurement is a value between 0 and 1 calculated by having a classifier predict the classes of the same data with which it was trained, and then computing the average loss rate (Marais, Patell and Wolfson 1984). The results are displayed in Table 6-1.

	SAM	AffectButton
Pleasure	.1888	.3116

Arousal	.2913	.4055

Table 6-1: The classification errors for the four tree classifiers.

For each annotation, we compared the categorical value with the mapped value from the SAM and the mapped value from the AffectButton. With *strict filtering*, we accepted only results where all these three values were the same. However, as seen in Table 6-1, the trees have considerable classification errors, so their mappings are not completely reliable. Thus, we considered a more lenient approach as well. With this *lenient filtering*, we accepted annotations for which least one of between AB and SAM values would map to the same category as that indicated in the categorical judgment, and the other was not the opposite of the other two.

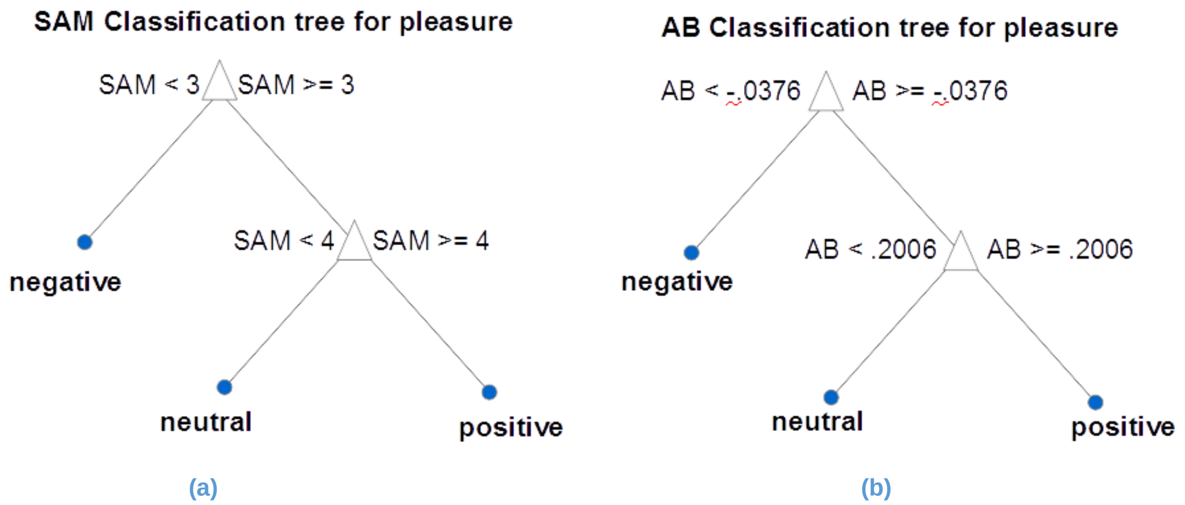


Figure 6-2: Visual representation of the tree classifier mapping the *pleasure* rating from the SAM to the categorical scale (a) and the classifier mapping the pleasure value from the AffectButton to the categorical scale (b). SAM (in a) is the acquired value from the SAM, measured from the most negative picture as 1 to the most positive picture as 5. AB (in b) is the acquired value from the AffectButton, measured as a real number from -1 to 1.

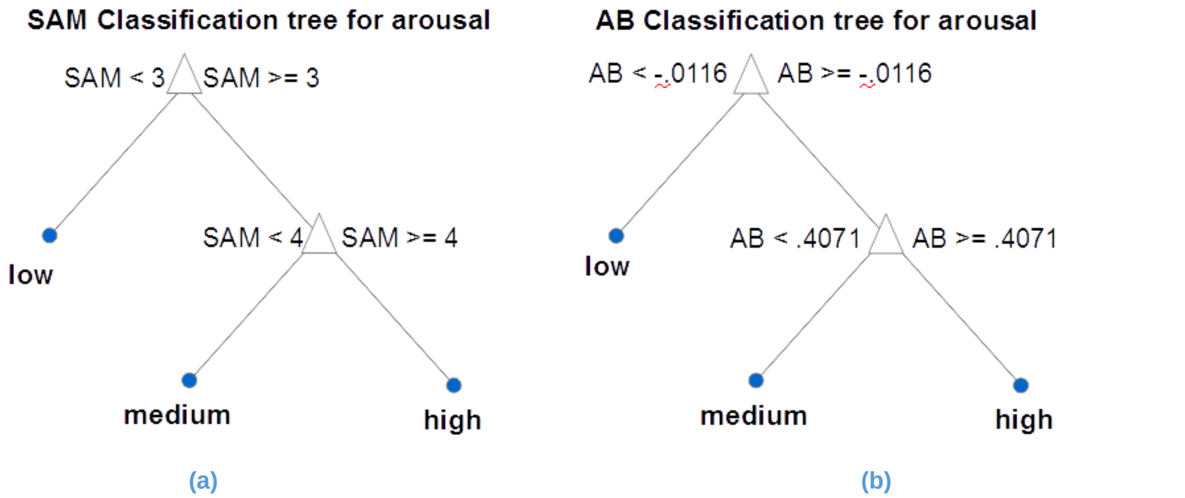


Figure 6-3: Visual representation of the tree classifier mapping the *arousal* rating from the SAM to the categorical scale (a) and the classifier mapping the value from the AffectButton to the categorical scale (b). SAM (in a) is the acquired value from the SAM, measured from the least aroused picture as 1 to the most aroused picture as 5. AB (in b) is the acquired value from the AffectButton, measured as a real number from -1 to 1.

6.2 Annotation Analysis

In this section, we explain the analysis we performed on the data. This analysis aims at addressing three core questions:

- How can videos be obtained that accurately portray depression and anxiety?
- How can the reliability of crowdsourced annotations be controlled effectively?

For the first question, we verify how accurately the videos portrayed the intended mood. For the second question, we analyse the effects of the mechanisms intended to control the reliability of the annotations.

Below we describe the analysis that was performed *for each filtering stage option*, to understand how each filter alters the answer to these questions. Ideally, the more (and stricter) filters applied, the more positive the answer to the first question should be, as unreliable ratings should be eliminated. To answer the second question we will evaluate both the rating reliability and the number of ratings remaining after each filtering stage, to understand better the trade-off between reliability and waste of resources (in terms of discarded, yet paid, annotations). The results are presented in Figure 6-4. Based on these results, an optimal set of filtering criteria will be chosen and discussed in Sections 6.3.

6.2.1 Comparing intended and perceived mood

In Chapter 4, we examined the relationship between the intended mood and the mood felt by the actors. In this chapter, we compare the intended mood (the one we induced in actors) and the mood perceived by the annotators. To facilitate comparison, we defined the intended mood in the same classes as the categorical values: low, medium and high. We define depression as low pleasure and low arousal, anxiety as low pleasure and high arousal and amusement as high pleasure and neutral arousal.

To verify whether the intended mood was also perceived as such, we compared the categorical description of the intended mood with the categorical annotations. As we also wished to make this comparison for the SAM and the AffectButton annotations, we used a tree classifier to map the ratings from the SAM and the AffectButton to the categories of the intended mood, as we did for the consistency filtering (see Section 5.2). For each filtering stage, we trained a tree classifier using the data remaining after applying those filtering options, so that it would not be trained with unreliable data. We then calculated the Spearman correlation between the intended mood and the annotated mood across all videos, taking each annotation as an individual point. For each layer of filtering, we calculated this correlation for all three tools. The results per filtering stage can be seen in the 'Spearman' columns in the tree nodes of Figure 6-4.

However, as there are only three ranks of affect in the correlated measurements, the Spearman correlation is a less accurate method for comparison than if there were more ranks. Thus, we use a second method to investigate the relation between intended and perceived mood. The annotation of the videos can be seen as a multi-class classification task, where the intended mood describes the expected classes and the perceived mood describes the predicted classes. With this view, we calculated three values from information retrieval: precision, recall and the F-score.

The *Precision* of a given class is a measure of the number of items 'correctly' assigned that class (as expected) divided by the total number of classified items (Basu, Hirsh and Cohen 1998). For example, the precision for 'positive' indicates how many videos that were perceived as positive were indeed intended to portray positives moods.

The *Recall* of a given class measures the number of items classified as expected divided by the number of items where that class was expected (Basu, Hirsh and Cohen 1998). For example, the recall for ‘low arousal’ indicates how many moods that were intended to portray low arousal were indeed perceived as having low arousal.

The *F-score* serves as a measure that combines precision and recall, calculated as the harmonic mean of the two (Goutte and Gaussier 2005). In the case of a multi-class classification problem such as this one, the mean F-score averaged over all classes is a measure of the effectiveness of a classifier (Sokolova and Lapalme 2009). Therefore, we use the mean F-score as an indication of how closely the annotated mood agrees with the intended mood. The results per filtering stage can be seen in the ‘F’ columns in the tree nodes of Figure 6-4.

6.2.2 Using crowdsourcing for annotation

The answer to the question whether crowdsourcing was appropriate to annotate affective databases consisted of two parts. Firstly, to evaluate the reliability of the annotations, we looked into measures of inter-rater reliability. This is typically measured using Cronbach’s alpha (Cronbach 1951). Cronbach’s alpha requires within subjects measurements. In our data, however, every annotator rated one video (or several, in different numbers depending on the annotator). Therefore, we used *Krippendorff’s alpha* (Hayes and Krippendorff 2007), which can be applied to data regardless of the amount of observers and sample size and can be used for incomplete or missing data (Y, et al. 2014).

Krippendorff’s alpha is calculated as follows:

$$\alpha = 1 - \frac{D_{observed}}{D_{expected}}$$

where $D_{observed}$ is the observed disagreement among the observed ratings, and $D_{expected}$ is the disagreement which is to be expected if the observed ratings are the result of chance rather than a reliable system (in this case, a consistent rating of mood). These disagreement measures are calculated using coincidence matrices. See (K. Krippendorff 2011) for more details.

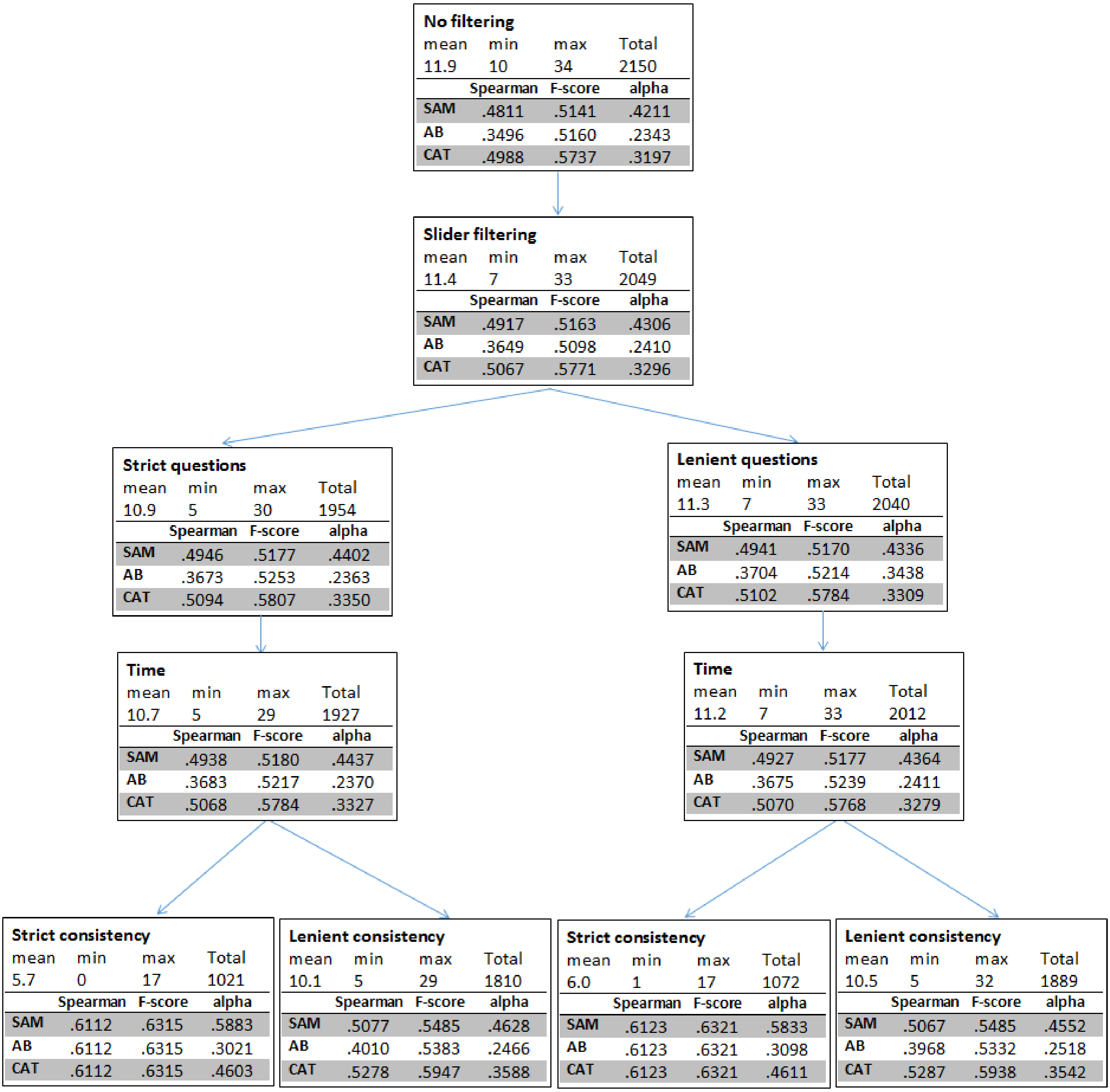
We calculated Krippendorff’s alpha for the ratings from each of the three tools. We used the ordinal scale for the SAM, the interval scale for the AffectButton and the nominal scale for the categorical value. See the results per filtering stage in the ‘alpha’ columns in Figure 6-4.

Second, we wanted to monitor the extent to which seeking reliability would result in a waste of resources. The more filtering criteria we applied, the more annotations we expected to see discarded. However, it was unclear whether discarding the annotations would result in severely pruning their amount, or whether instead they were already quite reliable to begin with, and thus their number would not be highly affected by the filtering. In addition, as we aimed to get at least six annotators per video, it is relevant to examine whether we reached that number after applying the filters. Thus, we examined the following general statistics in the data, after each filtering stage:

- The minimum, maximum and mean number of *annotations per video*, as well as the total number of annotations.
- The total *number of annotators* remaining.

These numbers are reported in the top row of each node of the filtering tree in Figure 6-4.

Filtering tree for pleasure



No filtering

mean	min	max	Total
11.9	10	34	2150

	Spearman	F-score	alpha
SAM	.4811	.5141	.4211
AB	.3496	.5160	.2343
CAT	.4988	.5737	.3197

Slider filtering

mean	min	max	Total
11.4	7	33	2049

	Spearman	F-score	alpha
SAM	.4917	.5163	.4306
AB	.3649	.5098	.2410
CAT	.5067	.5771	.3296

Strict questions

mean	min	max	Total
10.9	5	30	1954

	Spearman	F-score	alpha
SAM	.4946	.5177	.4402
AB	.3673	.5253	.2363
CAT	.5094	.5807	.3350

Lenient questions

mean	min	max	Total
11.3	7	33	2040

	Spearman	F-score	alpha
SAM	.4941	.5170	.4336
AB	.3704	.5214	.3438
CAT	.5102	.5784	.3309

Time

mean	min	max	Total
10.7	5	29	1927

	Spearman	F-score	alpha
SAM	.4938	.5180	.4437
AB	.3683	.5217	.2370
CAT	.5068	.5784	.3327

Time

mean	min	max	Total
11.2	7	33	2012

	Spearman	F-score	alpha
SAM	.4927	.5177	.4364
AB	.3675	.5239	.2411
CAT	.5070	.5768	.3279

Strict consistency

mean	min	max	Total
5.7	0	17	1021

	Spearman	F-score	alpha
SAM	.6112	.6315	.5883
AB	.6112	.6315	.3021
CAT	.6112	.6315	.4603

Lenient consistency

mean	min	max	Total
10.1	5	29	1810

	Spearman	F-score	alpha
SAM	.5077	.5485	.4628
AB	.4010	.5383	.2466
CAT	.5278	.5947	.3588

Strict consistency

mean	min	max	Total
6.0	1	17	1072

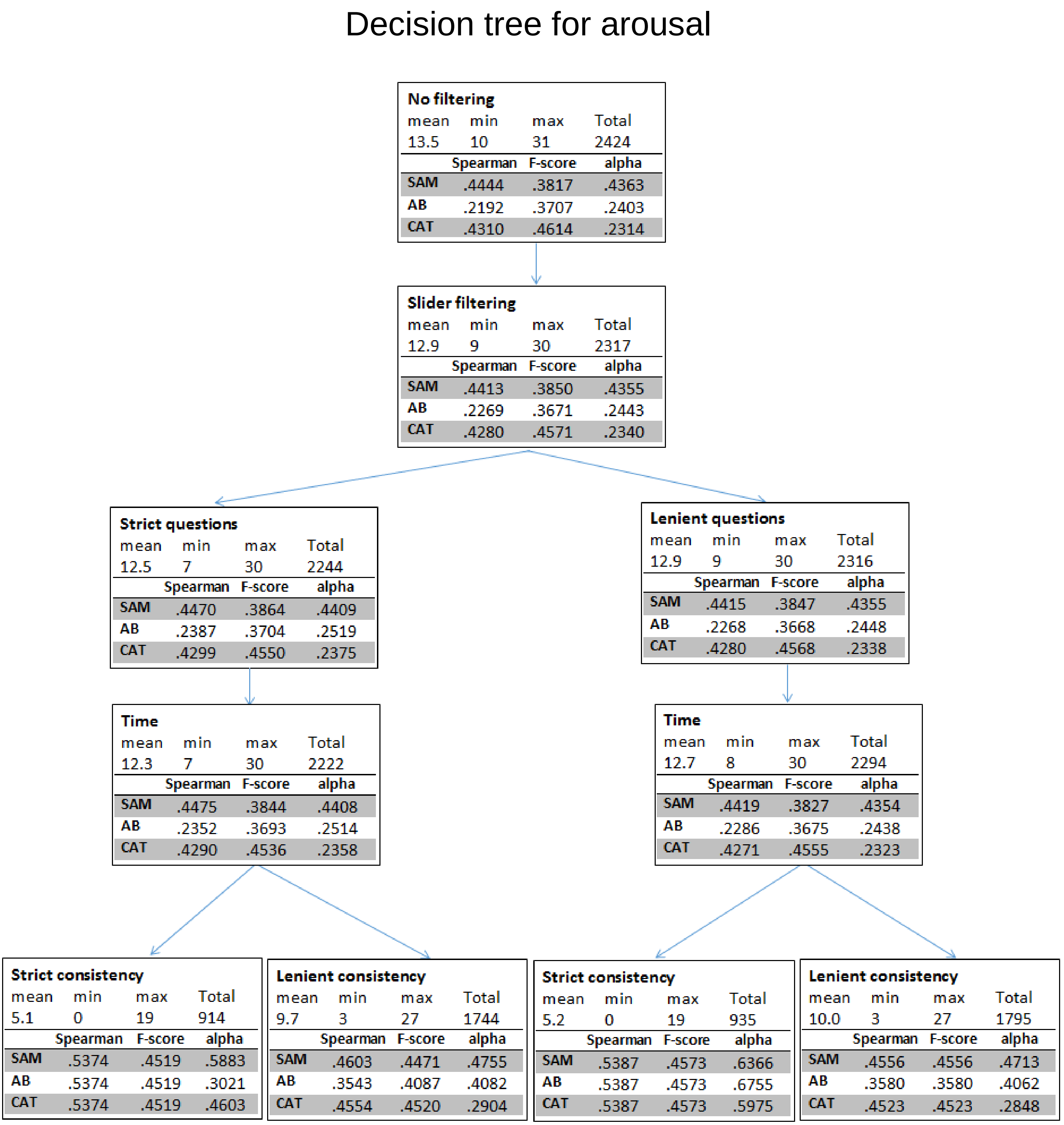
	Spearman	F-score	alpha
SAM	.6123	.6321	.5833
AB	.6123	.6321	.3098
CAT	.6123	.6321	.4611

Lenient consistency

mean	min	max	Total
10.5	5	32	1889

	Spearman	F-score	alpha
SAM	.5067	.5485	.4552
AB	.3968	.5332	.2518
CAT	.5287	.5938	.3542

(a)



Decision tree for arousal

No filtering

mean	min	max	Total
13.5	10	31	2424

	Spearman	F-score	alpha
SAM	.4444	.3817	.4363
AB	.2192	.3707	.2403
CAT	.4310	.4614	.2314

Slider filtering

mean	min	max	Total
12.9	9	30	2317

	Spearman	F-score	alpha
SAM	.4413	.3850	.4355
AB	.2269	.3671	.2443
CAT	.4280	.4571	.2340

Strict questions

mean	min	max	Total
12.5	7	30	2244

	Spearman	F-score	alpha
SAM	.4470	.3864	.4409
AB	.2387	.3704	.2519
CAT	.4299	.4550	.2375

Lenient questions

mean	min	max	Total
12.9	9	30	2316

	Spearman	F-score	alpha
SAM	.4415	.3847	.4355
AB	.2268	.3668	.2448
CAT	.4280	.4568	.2338

Time

mean	min	max	Total
12.3	7	30	2222

	Spearman	F-score	alpha
SAM	.4475	.3844	.4408
AB	.2352	.3693	.2514
CAT	.4290	.4536	.2358

Time

mean	min	max	Total
12.7	8	30	2294

	Spearman	F-score	alpha
SAM	.4419	.3827	.4354
AB	.2286	.3675	.2438
CAT	.4271	.4555	.2323

Strict consistency

mean	min	max	Total
5.1	0	19	914

	Spearman	F-score	alpha
SAM	.5374	.4519	.5883
AB	.5374	.4519	.3021
CAT	.5374	.4519	.4603

Lenient consistency

mean	min	max	Total
9.7	3	27	1744

	Spearman	F-score	alpha
SAM	.4603	.4471	.4755
AB	.3543	.4087	.4082
CAT	.4554	.4520	.2904

Strict consistency

mean	min	max	Total
5.2	0	19	935

	Spearman	F-score	alpha
SAM	.5387	.4573	.6366
AB	.5387	.4573	.6755
CAT	.5387	.4573	.5975

Lenient consistency

mean	min	max	Total
10.0	3	27	1795

	Spearman	F-score	alpha
SAM	.4556	.4556	.4713
AB	.3580	.3580	.4062
CAT	.4523	.4523	.2848

(b)

Figure 6-4: Filtering trees for pleasure (a) and arousal (b). Each node displays the statistics for the annotations that remain after applying that filter and the preceding filters.

6.2.3 Results

In this section, we comment on the findings presented in Figure 6-4 and evaluate the effects of the reliability control mechanisms. The results are presented in a tree, where each node represents a filtering stage. For each stage, we show the statistics of the annotations per video, the agreement with the intended mood, the average F-score over all three classes and the value of Krippendorff's alpha. We will discuss the effects of each filtering stage and conclude which set of filtering stages is most appropriate given our research questions.

6.2.3.1 Continuous data

The continuous data filtering removes 101 annotations for pleasure and 107 annotations for arousal. We note that for pleasure the minimum amount of annotations per video is seven. Upon closer inspection, we found there is one video with only seven annotations and there are five videos with eight. Four of these videos, including the one with seven remaining annotations, are shorter than one minute, which likely explains the lack of movement of the slider. Though the remaining two videos were longer, they were apparently annotated either by at least two annotators who either considered them uneventful, or were indeed not performing the task correctly.

We see a slight increase in alpha, F-score and correlation with intended mood for pleasure, suggesting that the filtering did indeed remove unreliable annotations. For arousal, the change is much smaller (<.01).

6.2.3.2 Content questions

When applying the lenient criterion, the content question filtering removes nine annotations for arousal and one for pleasure. This is as expected: it is very unlikely that an annotator would get more than one of these three questions wrong, unless they were randomly pressing buttons. From the small amount of removed annotations, we can conclude that the majority of the annotators were carefully watching the videos when annotating them.

After applying strict filtering, the minimum amount of annotations per video drops by two for both arousal and pleasure. For pleasure, one video had five annotators remaining. In this video, the actress was handed a plate and touched the food with her fork, but did not actually take a bite. It is likely the question of whether the person in the video was eating caused confusion in this video. For arousal, one video had seven annotators remaining. In this video, we could find no ambiguity: two annotators had simply answered the question incorrectly.

For both pleasure and arousal, the choice for a lenient or strict filtering appears to have little influence on the change in the alpha, F-score and correlation. For both the lenient and the strict path, these values increase slightly for pleasure. There is hardly any change for arousal.

6.2.3.3 Annotation time

The time filtering removed around twenty annotations for both arousal and pleasure, both after lenient and strict question filtering, and the effects are slight. Considering the leniency of our filtering condition, this is to be expected. The alpha, F-score and correlation slightly change, but with this small amount of removed annotations and such a small difference, we can hardly draw conclusions from this.

6.2.3.4 Consistency

The filtering stage with the most impact on the results is the consistency filtering. We see that applying strict consistency increases agreement with the intended mood and Krippendorff's alpha for all four ratings, compared to the lenient consistency, and in turn with the previous filtering stage. For example, in the arousal tree we see that after applying slider filtering, strict question filtering and time filtering, the Krippendorff's alpha for the SAM is 0.44. If we apply lenient consistency filtering, the alpha becomes 0.48. If we apply strict filtering instead, we obtain an alpha of 0.64. The same increase occurs in the alpha values of the other tools. We conclude that checking for consistency between annotation tools increases the reliability of the resulting annotations. For the F-score, this holds true for pleasure as well. The F-scores hardly change for any of the criteria for arousal, however.

The strict consistency filtering discards a very high number of annotations (over 45%): in other words, every two paid annotations, only one can be retained. It is also interesting to note that more annotations for arousal than for pleasure (59% for arousal as opposed to 47% for pleasure). A possible explanation is the way the AffectButton calculates the arousal value from the input. Whereas the pleasure and dominance values are directly calculated from the position of the cursor in the button, the arousal value is derived from the position. (Broekens and Brinkman 2013) This extra step in the calculation of the values may explain how the consistency is generally lower for arousal.

6.3 Optimal filtering path

The choice of criteria for the filters is a trade-off between reliability and amount of annotations.

The first three filtering stages affect the number of annotations and the results of the analyses far less than the consistency filtering. The main decision is whether to adopt the strict consistency method or the lenient one in the last stage. Whereas strict consistency filtering gives higher agreement and higher inter-rater reliability, it does leave 28 videos with less than three annotators. We stated in Chapter 5 that, based on (Cowie and McKeown 2010), we would want at least six annotators per videos in order to have a reliable set of ratings. Therefore, we consider the high amount of annotations that have to be excluded to fulfil the strict consistency criterion unacceptable. In light of this, *the path with strict question filtering and the lenient consistency filtering* is the second best option for both arousal and pleasure in terms of correlation with the intended mood and inter-rater reliability.

Following this path through the decision tree (i.e.: applying the continuous filtering, the strict questions filtering, the time filtering and the lenient consistency filtering) we find that we have to discard 16% of the annotations for pleasure and 28% for arousal. These percentages can be considered an indication of the excess annotations one needs to account for when planning a crowdsourcing campaign such as this one. In the remainder of this section, we will analyse the results after applying the chosen filtering.

6.3.1 Comparing intended and perceived mood

Taking the chosen path, we find that the Spearman correlation between the reported categorical mood and the intended mood is moderate for both pleasure ($r_s = .5278$, $p < .0001$) and arousal ($r_s = .4554$, $p < .0001$). We conclude that the intended mood and the perceived mood are indeed associated, although for a number of videos the perceived mood may differ from the intended one. On the positive side, this ensures diversity of portrayed moods in our database.

In addition to considering the F-score, we examine the confusion matrix for the annotation to reveal mismatches between the intended mood and perceived categorical mood. The confusion matrix for pleasure is displayed in Table 6-2 and the confusion matrix for arousal is shown in Table 6-4. Based on this confusion matrix, we calculated the precision, recall and F-score. The results are found below in Table 6-3 for pleasure and in Table 6-5 for arousal.

	Annotated Negative	Annotated Neutral	Annotated Positive
Expected Negative	877	484	44
Expected Neutral	61	133	50
Expected Positive	1	13	147

Table 6-2: Confusion Matrix for pleasure. Each column represents the instances in an annotated class while each row represents the instances in an expected class.

	Precision	Recall	F-score
Negative	.9340	.6242	.7483
Neutral	.2111	.5451	.3043
Positive	.6100	.9130	.7313
Average	.5850	.6941	.5947

Table 6-3: Precision, recall and F-score for pleasure, calculated from the confusion matrix in Table 6-2. The (bottom right) average F-score is the value presented in the decision tree for each filtering stage.

From Table 6-2, we see that out of 1405 expected negative moods, 484 were annotated as neutral (34%). This results in a low precision value for neutral, as seen in Table 6-3. Conversely, of the 244 expected neutral moods, 61 were annotated as negative (25%). Apparently, there was disagreement over the boundaries between negative and neutral mood. Aside from this mismatch, the expected and annotated moods coincide quite well, resulting in an acceptable average F-score.

	Annotated Low	Annotated Medium	Annotated High
Expected Low	387	140	13
Expected Medium	165	183	82
Expected High	156	390	228

Table 6-4: Confusion Matrix for arousal. Each column represents the instances in an annotated class while each row represents the instances in an expected class.

	Precision	Recall	F-score
Low	.5466	.7167	.6202
Medium	.2567	.4256	.3202
High	.7059	.2946	.4157
Average	.5031	.4789	.4520

Table 6-5: Precision, recall and F-score for pleasure, calculated from the confusion matrix in Table 6-4. The (bottom right) average F-score is the value presented in the decision tree for each filtering stage.

For arousal, we see in Table 6-4 that out of 430 moods expected to have medium arousal, 165 were attributed low arousal by the annotators (38%), out of the 774 moods expected to have high arousal, 390 were annotated as having medium arousal (50%). Apparently, in general a lower arousal was perceived than we expected. It is not surprising, then, that low arousal has the highest recall value of the three classes. This coincides with our findings in Section 5.5.1 where we concluded from the histograms that videos intended to portray anxiety (high arousal) was rated lower than expected on average. It is possible that, as the physical cues provided to the actors were the cues of elderly, their behaviour was perceived as less active than usual, even when in an anxious mood.

6.3.2 Inter-rater reliability

We now discuss the Krippendorff's alpha values for the annotations after applying the strictest filtering. For the reader's convenience, we single out these results from Figure 6-4 below in Table 6-6.

	Pleasure	Arousal
SAM	.46281	.47548
AffectButton	.2466	.4082
Categorical	.3588	.2904

Table 6-6: Table showing the values of Krippendorff's alpha for each dimension for each annotation tool after applying the strictest filtering.

In previous crowdsourcing experiments such as (Soleymani, Caro, et al. 2013) (Y. Baveye, et al. 2013), values for Krippendorff's alpha above 0.4 have been considered acceptable. Following this policy, we see that the alpha for the SAM values is acceptable for both pleasure and arousal. The alpha value for the ratings from the AffectButton is surprisingly low for pleasure, compared to the acceptable alpha for arousal. Further examining the decision tree for pleasure in Figure 6-4a, we see this value is relatively low no matter which filter is applied. It would be fruitful for future work to examine how the AffectButton is used differently for mood rating in crowdsourcing experiments compared to other tools. The alpha values for the categorical scale are rather low as well, compared to the values of the strict consistency. Apparently choosing lenient consistency over strict consistency implicates allowing more noise than desired in the categorical annotations. The SAM results suggest that our crowdsourcing experiment, after appropriate filtering, has produced reliable SAM annotations.

Chapter 7: Conclusions and Recommendations

In this chapter, we will summarize the work we have done and its results. We will discuss the implications of these results and finally provide recommendations for further research.

7.1 Contributions

This work concerned the creation of an annotated mood database to be used to train a mood recognition system in a care home for elderly, in particular to recognize two negative moods often occurring in care homes, namely depression and anxiety. This work is a part of the ACE project which is tasked with creating an affect adaptive system that detects the mood of a resident in a care home and adapts the lighting of the room to positively influence the resident's mood.

As the vast majority of existing affective databases are centred on emotion, this study contributes to the field by investigating how a mood database can be created and annotated. This task was performed in three stages. Firstly, we investigated how humans recognize mood in the same setting by interviewing caretakers in a care home. Specifically, we asked for the physical cues that the caretakers register when recognizing depression and anxiety. We also asked for the typical causes of these two moods. From these causes, we wrote a number of scenarios that represented the appropriate context for the videos which the database would comprise.

Secondly, we recorded visual data of the intended moods, as acted by professional actors who were also induced the negative moods they have to portray. The actors portrayed depression, anxiety, and additionally amusement as a counterexample. We instructed actors using the physical cues and scenarios from the interviews. The scenarios provided a suitable context to make the acting more believable, and the physical cues were incorporated in their acting using a theatre technique known as physical acting. We recorded these actors using a camera focussed on the face, a camera filming the whole body and a Kinect, so that we would acquire our moods in three modalities. To verify whether the induction was successful, we asked the actor to self-report his/her mood before the induction, after the induction, and after recording on a Self-Assessment Manikin.

Finally, the visual data was annotated, which is traditionally done by a small number of experts, a time consuming and expensive process. We used crowdsourcing as an alternative, allowing us to distribute the annotation task to a large group of online workers. Crowdsourcing implicates the risk of acquiring unreliable data. To investigate how we could increase the reliability of the annotations, we implemented a number of reliability control mechanisms. We then filtered the results of the campaign to remove unreliable annotations.

As a result, the contributions of this work can be summarized in three parts. We provide a new annotated mood database that can be used for mood recognition. We implemented an annotation tool specifically focussed on the annotation of videos in terms of both emotion and mood. We investigated the limitations, in terms of reliability, of using crowdsourcing for the affective annotation of videos.

7.2 Discussion of Results

We defined four research questions in Chapter 1. We will briefly discuss how we answered these questions and the implications of these answers.

RQ1. What are the defining features of the moods depression and anxiety when experienced by elderly?

Section 2.1 discusses the characteristics of mood and how to model them, providing insight in the nature of mood on a conceptual level. Section 4.1 presents the results of a series of interviews with caretakers in a care centre, containing a list of cues and contexts that are associated with depression and anxiety for elderly. These results provide insights in the nature of the moods on a behavioural level, which is necessary to create a representative setting for the acted material.

RQ2. How can media be obtained that accurately portray the intended mood?

From the SAMs filled out by the actors, a Wilcoxon Signed-rank test revealed that the mood reported by the actors was affected significantly by our mood induction procedure, and the change of the mood was as intended. See Section 4.4.2 for details. Furthermore, as shown in Section 6.3.1 the annotated mood was moderately correlated with the intended mood for both pleasure ($r_s = .5278$, $p < .0001$) and arousal ($r_s = .4554$, $p < .0001$). These were supported by the calculation of confusion matrices, from which an average F-score of .5947 was calculated for pleasure and .4520 for arousal. From these facts, we conclude that the intended, felt and perceived mood were consistent. Thus, we have shown how actors and mood induction can be combined to capture video material depicting the desired affective content.

RQ3. How can the reliability of the resulting annotations be controlled effectively?

We developed a tool for mood annotation and implemented reliability control mechanisms, described in Chapter 5. We measured reliability as the agreement between raters by calculating Krippendorff's alpha. In Section 6.2 we have demonstrated how the reliability control mechanisms can be used to remove unreliable annotations with different levels of strictness and how this process increases the reliability of the resulting annotations. We saw small increases in Krippendorff's alpha, around .01 for each stage, when filtering the annotations of pleasure according to the first three filtering stages. These stages were based on continuous data, answers to content-questions and task completion time. For arousal, the alpha changed very little upon applying these filtering stages, implying arousal reliability is more difficult to control. The consistency filtering had a very large effect on the data, particularly when applying the strictest filtering for which allowed for only annotations that were consistent over all three annotation tools. This filtering increased the alpha from .4 to .6 for both arousal and pleasure, independent of the filtering criteria applied beforehand. From this, we conclude that workers who were more consistent throughout their own annotation agreed more with other annotators.

RQ4. Is crowdsourcing an appropriate tool to obtain accurate annotations for a mood database?

After using reliability control mechanisms to filter out unreliable annotations in Section 6.2, we performed analyses on the resulting data. Section 6.3.2 presents the analysis of the reliability of our annotations using Krippendorff's alpha. We found acceptable values of alpha for the SAM annotations: .463 for pleasure and .475 for arousal. This work shows that crowdsourcing can indeed be used as a tool for affective annotation of videos. It is worth noting, however, that these values of alpha are considerably lower for the AffectButton and the categorical mood. While the alpha for arousal for the AffectButton was acceptable (.408), the alpha for pleasure was low (.247). As the AffectButton has been validated with the SAM (Broekens, Joost Broekens 2015), we would have expected the alphas for the AffectButton to be similar for the SAM. It is interesting to note that when comparing intended and perceived mood, the correlations and F-scores for arousal were mostly lower than those for pleasure. In conclusion, while annotations of pleasure were more in agreement with the expected moods, the annotations of arousal were more in agreement with those of other annotators. It is worthwhile to investigate these findings in future work, and whether they stem from a difference in interpretation of the dimensions, a difference in the measuring tools, or a difference in the videos themselves.

7.3 Limitations and Recommendations

This section addresses the limitations of the studies, and presents recommendations for research to be conducted in the future.

It is worthwhile to further investigate the step we make from the interviews we have held to the instructions we gave to our actors. Although we have established physical cues and scenarios that caretakers associate with each mood on a conscious level, much of human emotion recognition happens on a subconscious level (Dimberg, Thunberg and Elmehed 2000). We filmed the caretakers imitating an elderly in a certain mood to register some of the features that they might imitate unconsciously. But this subject could be explored in much greater depth. If the actors and their director spent some time with the elderly as well, and specifically those with depression and anxiety, the director could create more accurate scenarios from first-hand experiences. The actors could imitate the elderly from a direct experience. A caretaker could attend the recordings and give feedback on how realistic they find the situation and acting. We believe that this could lead to more realistic visual data that would be more applicable for a mood recognition system for elderly. This is a complicated solution, however, as it involves many privacy issues.

To expand the number of applications of the mood database and provide more insight in mood recognition, it would be useful to add more moods to the database, e.g. moods that cover parts of the pleasure-arousal domain not currently covered (i.e. low arousal and high pleasure). In particular, while we recorded and annotated videos of the mood amusement, the database would be more complete if it would be researched to the same extent as the other two moods by performing an interview with caretakers about recognition and causes of amusement and recording SAMs of it.

In this research, we described and annotated the mood in terms of pleasure and arousal. As stated in Section 2.1.2.1 the third dimension originally included in the dimensional model is dominance (Russell and Mehrabian 1977). It would be interesting to use dominance for annotation as well. This would provide a more precise description of the videos, and allow us to study in practice whether the mood recognition system improved upon adding this dimension to the input data for training and testing.

In the annotation phase, the face videos were annotated. Due to time restrictions, the body videos are not yet been provided with annotations. An obvious next step is to annotate the body videos, and compare the results. As it has been suggested that moods are expressed via the body (Parkinson, et al. 1996), the annotated moods may be closer to the intended moods when seeing the whole body. Therefore, the recognition system may perform better when trained using the features from the body.

A limitation of the crowdsourcing setup was that the videos had to be cut up into parts to prevent the task from becoming too long. It is interesting to find out whether this impacted how the mood was perceived, and find out whether parts of a video were annotated similarly.

The videos were annotated in terms of mood as well as emotions. An interesting new research goal would be to explore the relationship between the two. We stated in Chapter 2 that we recognize a mood by the emotions we associate with it (P. Ekman 1999). Then it is a reasonable hypothesis that there are relationships between the annotated mood and emotions. It would also be interesting to see which moods arise when we induce emotions. An example would be to have one of the existing emotion databases discussed in Chapter 2 annotated in terms of mood and investigate how these mood annotations relate to the already existing emotion annotations. This could provide useful new insights for the topic of mood recognition.

In conclusion, this study has contributed a mood database that can be used for the automatic recognition of moods. We close with the hope that this study will provide one more step on the road towards the creation of systems that can act upon the true affective state of a human being.

Bibliography

Abadi, Mojtaba Khomami, et al. “A Multi-task Learning Framework for Time-continuous Emotion Estimation from Crowd Annotations.” *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia.* 2014.

Albersnagel, Frans A. “Velten and musical mood induction procedures: A comparison with accessibility of thought associations.” *Behaviour Research and Therapy* 26, no. 1 (1988): 79-95.

Artstein, Ron, and Massimo Poesio. “Inter-coder agreement for computational linguistics.” *Computational Linguistics* 34, no. 4 (2008): 555-596.

B., Fasel, and Luettin J. “Automatic facial expression analysis: a survey.” *Pattern recognition* 36, no. 1 (January 2003): 259-75.

Bänziger, Tanja, and Klaus Scherer. “Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus.” *Affective computing and intelligent interaction*, 2007: 476-487.

Barraclough, Jennifer. “ABC of palliative care: depression, anxiety, and confusion.” *BMJ* 315, no. 7119 (1997): 1365-1368.

Bartlett, Marian Stewart, Gwen Littlewort, Claudia Lainscsek, Ian Fasel, and Javier Movellan. “Machine learning methods for fully automatic recognition of facial expressions and facial actions.” *Systems, Man and Cybernetics, 2004 IEEE International Conference on.* IEEE, 2004. 592-597.

Basu, Chumki, Haym Hirsh, and William Cohen. “Recommendation as classification: Using social and content-based information in recommendation.” *Aaai/iaai.* 1998. 714-720.

Batliner, Anton, et al. “"You Stupid Tin Box"-Children Interacting with the AIBO Robot: A Cross-linguistic Emotional Speech Corpus.” *LREC.* Lisbon, 2004.

Baveye, Y., J.-N. Bettinelli, E. Dellandrea, L. Chen, and C. Chamare. “A Large Video Database for Computational Models of Induced Emotion.” *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on.* Geneva, 2013. 13-18.

Baveye, Yoann, Jean-Noël Bettinelli, Emmanuel Dellandréa, Liming Chen, and Christel Chamaret. “A large video database for computational models of induced emotion.” *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on.* 2013. 13-18.

Beedie, Christopher, Peter Terry, and Andrew Lane. “Distinctions between emotion and mood.” *Cognition and Emotion* 19, no. 6 (2005): 847-878.

Behrend, Tara S., David J. Sharek, Adam W. Meade, and Eric N. Wiebe. “The viability of crowdsourcing for survey research.” *Behavior research methods* 43, no. 3 (2011): 800-813.

Berking, Matthias, and Brian Whitley. *Affect Regulation Training: A Practitioners' Manual.* New York: Springer, 2014.

Brabham, Daren C. “Crowdsourcing as a model for problem solving an introduction and cases.” *Convergence: the international journal of research into new media technologies* 14, no. 1 (2008): 75-90.

Bradley, Margaret M., and Peter J. Lang. “Measuring emotion: the self-assessment manikin and the semantic differential.” *Journal of behavior therapy and experimental psychiatry* 25, no. 1 (March 1994): 49-59.

Broekens, Joost. July 2015. http://www.joostbroekens.com/.

Broekens, Joost, and Willem-Paul Brinkman. “AffectButton: a method for reliable and valid affective self-report.” *International Journal of Human-Computer Studies* 71, no. 6 (2013): 641-667.

Buchwald, Alexander M., Stephen Strack, and James C. Coyne. “Demand characteristics and the Velten mood induction procedure.” *Journal of Consulting and Clinical Psychology* 49, no. 3 (1981): 478.

Busso, C., and S. Narayanan. “Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database.” *INTERSPEECH.* Brisbane, Australia, 2008. 1670-1673.

Busso, Carlos, et al. “IEMOCAP: Interactive emotional dyadic motion capture database.” *Language Resources and Evaluation* 42, no. 4 (November 2008): 335-359.

Calvo, Rafael A., Sidney D'Mello, Jonathan Gratch, and Arvid Kappas. *The Oxford handbook of affective computing.* USA: Oxford University Press, 2014.

Cannon, WB. “The James-Lange theory of emotions: A critical examination and an alternative theory.” *The American journal of psychology* 100, no. 3/4 (1987): 567-586.

Caridakis, G., J. Wagner, A. Raouzaiou, Z. Curto, E. André, and K. Karpouzis. “A multimodal corpus for gesture expressivity analysis.” *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 2010: 80.

Cole, Toby. *Acting: a handbook of the Stanislavski method.* Crown Trade Paperbacks, 1995.

Cowie, Roddy, and Gary McKeown. “Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme.” Report, 2010.

Cowie, Roddy, and Martin Sawey. “GTrace - General trace program from Queen’s, Belfast.” 2011.

Cowie, Roddy, and Randolph R. Cornelius. “Describing the emotional states that are expressed in speech.” *Speech communication* 40, no. 1 (2003): 5-32.

Cowie, Roddy, Ellen Douglas-Cowie, and Cate Cox. “Beyond emotion archetypes: Databases for emotion modelling using neural networks.” *Neural networks* 18, no. 4 (2005): 371-388.

Cowie, Roddy, Ellen Douglas-Cowie, and Susie Savvidou. “'FEELTRACE': An instrument for recording perceived emotion in real time.” *ISCA tutorial and research workshop (ITRW) on speech and emotion.* 2000.

Cronbach, L. J. “Coefficient alpha and the internal structure of tests.” *Psychometrika* 16, no. 3 (1951): 297-334.

CrowdFlower. *CML (CrowdFlower Markup Language) Overview.* 2015. https://success.crowdflower.com/hc/en-us/articles/202817989-CML-and-Instructions-CML-CrowdFlower-Markup-Language-Overview (accessed June 29, 2016).

Davidson, R. J. “On emotion, mood and related affective constructs.” In *The nature of emotion*, 51-55. Oxford, UK: Oxford University Press, 1994.

De la Torre, Fernando, and Jeffrey F. Cohn. “Facial expression analysis.” In *Visual analysis of humans*, by Thomas B. Moeslund, Adrian Hilton, Volker Krüger and Leonid Sigal, 377-409. London: Springer, 2011.

Devillers, Laurence, Laurence Vidrascu, and Lori Lamel. “Challenges in real-life emotion annotation and machine learning based detection.” *Neural Networks* 18, no. 4 (2005): 407-422.

Dimberg, Ulf, Monika Thunberg, and Kurt Elmehed. “Unconscious facial reactions to emotional facial expressions.” *Psychological science 11, no. 1*, 2000: 86-89.

D'mello, Sidney K., and Jacqueline Kory. “A review and meta-analysis of multimodal affect detection systems.” *ACM Computing Surveys (CSUR)* 47, no. 3 (2015): 43.

Doan, Anhai, Raghu Ramakrishnan, and Alon Y. Halevy. “Crowdsourcing systems on the world-wide web.” *Communications of the ACM* 54, no. 4 (2011): 86-96.

Douglas-Cowie, E., et al. “The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data.” Edited by Ana C. R. Paiva et al. *International Conference on Affective Computing and Intelligent Interaction.* Springer Berlin Heidelberg, 2007. 488-500.

Douglas-Cowie, Ellen, Nick Campbell, Roddy Cowie, and Peter Roach. “Emotional speech: Towards a new generation of databases.” *Speech communication* 40, no. 1 (2003): 33-60.

Downs, Julie S., Mandy B. Holbrook, Steve Sheng, and Lorrie Faith Cranor. “Are your participants gaming the system?: screening mechanical turk workers.” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2010. 2399-2402.

Duric, Zoran, et al. “Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction.” *Proceedings of the IEEE* 90, no. 7 (2002): 1272-1289.

E, Velten. “A laboratory task for induction of mood states.” *Behaviour research and therapy* 6, no. 4 (1968): 473-482.

E., Hudlicka. “To feel or not to feel: The role of affect in human–computer interaction.” *International journal of human-computer studies* 59, no. 1 (July 2003): 1-32.

Eickhoff, Carsten, and Arjen P. de Vries. “Increasing cheat robustness of crowdsourcing tasks.” *Information Retrieval* 16, no. 2 (2013): 121-137.

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

Ekman, Paul. “Basic emotions.” *Handbook of cognition and emotion* 98 (1999): 45-60.

Elfenbein, Hillary Anger, and Nalini Ambady. “On the universality and cultural specificity of emotion recognition: a meta-analysis.” *Psychological bulletin* 128, no. 2 (2002): 203.

Fessl, Angela, Verónica Rivera-Pelayo, Viktoria Pammer, and Simone Braun. “Mood tracking in virtual meetings.” In *21st century learning for 21st century skills*, edited by Andrew Ravenscroft, Stefanie Lindstaedt, Carlos Delgado Kloos and Davinia Hernández-Leo, 377-382. Springer Berlin Heidelberg, 2012.

Forbes-Riley, Katherine, and Diane J. Litman. “Adapting to Student Uncertainty Improves Tutoring Dialogues.” In *Artificial Intelligence in Education*, by Vania Dimitrova, 33-40. IOS Press, 2009.

G, Castellano, Caridakis G, Camurri A, Karpouzis K, Volpe G, and Kollias S. “Body gesture and facial expression analysis for automatic affect recognition.” In *Blueprint for affective computing: A sourcebook*, by Tanja Bänziger, Etienne Roesch Klaus R. Scherer, 245-255. 2010.

Gadiraju, U., R. Kawase, and S. Dietze. “A taxonomy of microtasks on the web.” *Proceedings of the 25th ACM conference on Hypertext and social media.* 2014. 218-223.

Goutte, Cyril, and Eric Gaussier. “A probabilistic interpretation of precision, recall and F-score, with implication for evaluation.” *European Conference on Information Retrieval.* Springer Berlin Heidelberg, 2005. 345-359.

Grimm, M. “The Vera am Mittag German audio-visual emotional speech database.” *Multimedia and Expo, 2008 IEEE International Conference on. IEEE.* 2008. 865 - 868.

Grimm, M., K. Kroschel, E. Mower, and S. Narayanan. “Primitives-based Evaluation and Estimation of Emotions in Speech.” *Speech Communication* 49, no. 10 (2007): 787-800.

Gross, James J. “Emotion regulation: Conceptual and empirical foundations.” In *Handbook of emotion regulation* , by James J Gross, 3-20. New York: NY: Guilford Press, 2014.

Gross, James J., and Robert W. Levenson. “Emotion elicitation using films.” *Cognition & Emotion* 9, no. 1 (1995): 87-108.

Gross, M. M., E. A. Crane, and B. L. Fredrickson. “Methodology for assessing bodily expression of emotion.” *Journal of Nonverbal Behavior* 34, no. 4 (2010): 223-248.

Gunes, H., and M. Piccardi. “A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior.” *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* 1 (2006): 1148 - 1153.

Gunes, Hatice, and Massimo Piccardi. “Automatic temporal segment detection and affect recognition from face and body display.” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, no. 1 (2009): 64-84.

Gunes, Hatice, Massimo Piccardi, and Maja Pantic. “From the lab to the real world: Affect recognition using multiple cues and modalities.” In *Affective computing: focus on emotion expression, synthesis, and recognition*, 185-218. Tech Education and Publishing, 2008.

Gwet, Kilem Li. “Computing inter-rater reliability and its variance in the presence of high agreement.” *British Journal of Mathematical and Statistical Psychology* 61, no. 1 (2008): 29-48.

Hayes, Andrew F., and Klaus Krippendorff. “Answering the call for a standard reliability measure for coding data.” *Communication methods and measures* 1, no. 1 (2007): 77-89.

Hirth, Matthias, Tobias Hoßfeld, and Phuoc Tran-Gia. “Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms.” *Mathematical and Computer Modelling* 57, no. 11 (2013): 2918-2932.

—. “Anatomy of a crowdsourcing platform-using the example of microworkers.com.” *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on.* IEEE, 2011. 322-329.

Hoßfeld, Tobias, et al. “Best Practices and Recommendations for Crowdsourced QoE-Lessons learned from the Qualinet Task Force "Crowdsourcing".” 2014.

Hoßfeld, Tobias, et al. “Best practices for QoE crowdtesting: QoE assessment with crowdsourcing.” *IEEE Transactions on Multimedia* 16, no. 2 (2014): 541-558.

Hoßfeld, Tobias, Michael Seufert, Matthias Hirth, Thomas Zinner, Phuoc Tran-Gia, and Raimund Schatz. “Quantification of YouTube QoE via crowdsourcing.” *Multimedia (ISM), 2011 IEEE International Symposium on.* Dana Point CA, 2011. 494-499.

Howe, Jeff. *Crowdsourcing: Tracking the Rise of the Amateur.* June 2006. http://crowdsourcing.typepad.com/cs/2006/06/.

Howe, Jeff. “The rise of crowdsourcing.” *Wired magazine* 14, no. 6 (2006): 1-4.

Hsueh, Pei-Yun, Prem Melville, and Vikas Sindhwani. “Data quality from crowdsourcing: a study of annotation selection criteria.” *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*, 2009.

Hu, Xiao. “Music and Mood: Where Theory and Reality Meet.” 2010.

Hudlicka, Eva. “To feel or not to feel: The role of affect in human–computer interaction.” *International journal of human-computer studies* 59, no. 1 (2003): 1-32.

I., Cohen, Sebe N., Garg A., Chen LS., and Huang TS. “Facial expression recognition from video sequences: temporal and static modeling.” *Computer Vision and image understanding* 91, no. 1 (August 2003): 160-187.

Jacko, Julie A. *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications.* CRC press, 2012.

Jana, Abhijit. *Kinect for Windows SDK Programming Guide.* Packt Publishing Ltd, 2012.

Katsimerou, Christina, Judith A. Redi, and Ingrid Heynderickx. “A computational model for mood recognition.” *International Conference on User Modeling, Adaptation, and Personalization.* Springer International Publishing, 2014. 122-133.

Kenealy, P. “Validation of a music mood induction procedure: Some preliminary findings.” *Cognition & Emotion* 2, no. 1 (1988): 41-48.

Ketai, Richard. “Affect, mood, emotion, and feeling: semantic considerations.” *The American journal of psychiatry* 132, no. 11 (November 1975): 1215-1217.

Kittur, Aniket, Ed H. Chi, and Bongwon Suh. “Crowdsourcing user studies with Mechanical Turk.” *Proceedings of the SIGCHI conference on human factors in computing systems.* 2008.

Klaus R. Scherer, Tanja Bänziger, Etienne Roesch. *A Blueprint for Affective Computing.* Oxford: Oxford University Press, 2010.

Kleinsmith, Andrea, and Nadia Bianchi-Berthouze. “Affective body expression perception and recognition: A survey.” *IEEE Transactions on Affective Computing* 4, no. 1 (2013): 15-33.

Krippendorff, K. *Computing Krippendorff's Alpha-Reliability.* 2011. http://repository.upenn.edu/asc_papers/43.

Krippendorff, Klaus. “Content analysis: An introduction to its methodology.” In *International encyclopedia of communication*, by G. Gerbner, W. Schramm, T. L. Worth, & L. Gross E. Barnouw, 403-407. New York: Oxford, 1989.

Kuijsters, Andre, Judith Redi, Boris de Ruyter, and Ingrid Heynderickx. “Lighting to Make You Feel Better: Improving the Mood of Elderly People with Affective Ambiences.” *PloS one* 10, no. 7 (2015).

Kulkarni, Saket S., Narender P. Reddy, and S. I. Hariharan. “Facial expression (mood) recognition from facial images using committee neural networks.” *Biomedical engineering online* 8, no. 1 (2009): 1-12.

Lane, Andrew M. and Terry, Peter C. “The nature of mood: Development of a conceptual model with a focus on depression.” *Journal of Applied Sport Psychology* 12, no. 1 (2000): 16-33.

Lee, Diana TF, Jean Woo, and Ann E Mackenzie. “A review of older people’s experiences with residential care placement.” *Journal of advanced nursing* 37, no. 1 (2002): 19-27.

Liping Shen, Minjuan Wang and Ruimin Shen. “Affective e-Learning: Using “Emotional” Data to Improve Learning in Pervasive Learning Environment.” *Journal of Educational Technology & Society* 12, no. 2 (April 2009): 176-189.

Ltd, Happyworm. January 2015. http://jplayer.org/.

Marais, M. Laurentius, James M. Patell, and Mark A. Wolfson. “The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications.” *Journal of accounting Research* 22 (1984): 87-114.

Marsella, Stacy, Jonathan Gratch, and Paolo Petta. “Computational models of emotion.” *A Blueprint for Affective Computing-A sourcebook and manual 11, no. 1*, 2010: 21-46.

McDuff, D. J., R. E. Kaliouby, and R. W. Picard. “Crowdsourcing Facial Responses to Online Videos.” *Affective Computing, IEEE Transactions on* 3, no. 4 (2012): 456-468.

McKeown, Gary, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. “The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent.” *Affective Computing, IEEE Transactions on* 3, no. 1 (April 2012): 5-17.

Mehrabian, A., and J. A. Russell. *An approach to environmental psychology.* Cambridge: the MIT Press, 1974.

Metallinou, Angeliki, and Shrikanth Narayanan. “Annotation and processing of continuous emotional attributes: Challenges and opportunities.” *Automatic Face and Gesture*

Recognition (FG), 2013 10th IEEE International Conference and Workshops on. Shanghai, 2013. 1-8.

Metallinou, Angeliki, et al. “The USC CreativeIT database: a multimodal database of theatrical improvisation.” *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality 18 May 2010* 55 (May 2010).

Mitra, Sushmita, and Tinku Acharya. “Gesture recognition: A survey.” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 37, no. 3 (2007): 311-324.

Morris, Robert R., and Daniel McDuff. “Crowdsourcing Techniques for Affective Computing.” In *The Oxford Handbook of Affective Computing (2014): 384.*, 384-394. 2014.

Morris, W.M. “A functional analysis of the role of mood.” *Review of personality and social psychology* 13 (1992): 256-293.

Nay, Rhonda. “Nursing home residents' perceptions of relocation.” *Journal of Clinical Nursing* 4, no. 5 (1995): 319-325.

Nowak, Stefanie, and Stefan Rüger. “How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation.” *Proceedings of the international conference on Multimedia information retrieval.* 2010.

Pantic, Maja, and Leon J. M. Rothkrantz. “Automatic analysis of facial expressions: The state of the art.” *IEEE Transactions on pattern analysis and machine intelligence* 22, no. 12 (2000): 1424-1445.

Parkinson, B., P. Totterdell, R. B. Briner, and S. Reynolds. *Changing moods: The psychology of mood and mood regulation.* Harlow, UK: Addison Wesley Longman, 1996.

Picard, Rosalind W. “Toward computers that recognize and respond to user emotion.” *IBM systems journal* 39, no. 3.4 (2000): 705-719.

Picard, Rosalind W., Elias Vyzas, and Jennifer Healey. “Toward machine emotional intelligence: Analysis of affective physiological state.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, no. 10 (2001): 1175-1191.

Pignatiello, Michael F., Cameron J. Camp and Lee A. Rasar. “Musical mood induction: An alternative to the Velten technique.” *Journal of Abnormal Psychology* 95, no. 3 (1986): 295-297.

Ramos, Carlos, Juan Carlos Augusto, and Daniel Shapiro. “Ambient intelligence—the next step for artificial intelligence.” *Intelligent Systems* 23, no. 2 (2008): 15-18.

Randolph, J. “Free-Marginal Multirater Kappa (multirater κ free): An Alternative to Fleiss’ Fixed-Marginal Multirater Kappa.” *Joensuu Learning and Instruction Symposium.* 2005.

Reeves, Byron, and Clifford Nass. *How people treat computers, television, and new media like real people and places.* CSLI Publications and Cambridge university press, 1996.

Riek, Laurel D., Maria F. O’connor, and Peter Robinson. “Guess what? a game for affective annotation of video using crowd sourcing.” In *Affective computing and intelligent interaction*, 277-285. Springer Berlin Heidelberg, 2011.

Roddy Cowie, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, Paul Stapleton. "Gtrace: General trace program compatible with EmotionML." *Humaine Association Conference on Affective Computing and Intelligent Interaction.* 2013. 709-710.

Rokach, Lior, and Oded Maimon. *Data mining with decision trees: theory and applications.* World scientific, 2014.

Rottenberg, Jonathan. "Mood and emotion in major depression." *Current Directions in Psychological Science* 14, no. 3 (2005): 167-170.

Rottenberg, Jonathan, James J. Gross, and Ian H. Gotlib. "Emotion context insensitivity in major depressive disorder." *Journal of abnormal psychology* 114, no. 4 (2005): 627-639.

Russell, James A. "Core affect and the psychological construction of emotion." *Psychological Review* 110, no. 1 (January 2003): 145-172.

Russell, James A., and Albert Mehrabian. "Evidence for a three-factor theory of emotions." *Journal of research in Personality* 11, no. 3 (1977): 273-294.

Russell, James A., and L. Feldman-Barrett. "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant." *Journal of Personality and Social Psychology* 76 (1999): 805-819.

Sheng, Victor, Foster Provost, and G. Panagiotis Ipeirotis. "Get another label? Improving data quality and data mining using multiple, noisy labelers." *Proceeding of KDD 2008.* 2008. 614–622.

Siegert, Ingo, Ronald Böck, and Andreas Wendemuth. "Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements." *Journal on Multimodal User Interfaces* 8, no. 1 (2014): 17-28.

Sneddon, Ian, Margaret McRorie, Gary McKeown, and Jennifer Hanratty. "The belfast induced natural emotion database." *Affective Computing, IEEE Transactions on* 3, no. 1 (April 2012): 32-41.

Snel, John, Alexey Tarasov, Charlie Cullen, and Sarah Jane Delany. "A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpora." *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals.* Istanbul, 2012.

Snow, Rion, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. "Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks." *Proceedings of the conference on empirical methods in natural language processing.* 2008.

Sokolova, Marina, and Guy Lapalme. "A systematic analysis of performance measures for classification tasks." *Information Processing & Management* 45, no. 4 (2009): 427-437.

Soleymani, Mohammad, Maja Pantic, and Thierry Pun. "Multimodal emotion recognition in response to videos." *Affective Computing, IEEE Transactions on* 3, no. 2 (2012): 211-223.

Soleymani, Mohammad, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. "1000 songs for emotional analysis of music." *CrowdMM '13 Proceedings of* *the 2nd ACM international workshop on Crowdsourcing for multimedia.* New York, 2013. 1-6.

Stanislavski, Constantin. *Building A Character.* Theatre Arts Books, 1989.

Surowiecki, J. *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations.* New York: Doubleday, 2004.

Sutherland, G., B. Newman, and S. Rachman. "Experimental investigations of the relations between mood and intrusive unwanted cognitions." *British Journal of Medical Psychology* 55, no. 2 (1982): 127-138.

Szwoch, Mariusz. "FEEDB: a multimodal database of facial expressions and emotions." *Human System Interaction (HSI), 2013 The 6th International Conference on* The 6th International Conference on (June 2013): 524 - 531.

Tao, Jianhua, and Tieniu Tan. "Affective computing: A review." In *Affective computing and intelligent interaction*, edited by Rosalind Picard, 981-995. Springer Berlin Heidelberg, 2005.

Tarasov, A., C. Cullen, and S. Delany. "Using Crowdsourcing for labeling emotional speech assets." *W3C workshop on Emotion ML.* Paris, 2010.

Thayer, R. E. *The origin of everyday moods.* Oxford, UK: Oxford University Press, 1996.

Varshney, Upkar. "Pervasive healthcare and wireless health monitoring." *Mobile Networks and Applications* 12, no. 2-3 (2007): 113-127.

Västfjäll, Daniel. "Emotion induction through music: A review of the musical mood induction procedure." *Musicae Scientiae* 5, no. 1 suppl (2002): 173-211.

Velten, Emmett. "A laboratory task for induction of mood states." *Behaviour research and therapy* 6, no. 4 (1968): 473-482.

Vidrascu, L., and L. Devillers. "Annotation and Detection of Blended Emotions in Real Human-Human Dialogs Recorded in a Call Center." *Proc. of ICME 2005.* Orsay, France, 2005. 944–947.

Wang, Rui. *Augmented Reality with Kinect.* Packt Publishing Ltd, 2013.

Watson, D., L. A. Clark, and A. Tellegen. "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales." *Journal of Personality and Social Psychology* 54, no. 6 (1988): 1063-1070.

Westermann, Rainer, G. Stahl, and F. Hesse. "Relative effectiveness and validity of mood induction procedures: A meta-analysis." *European Journal of Social Psychology* 26, no. 4 (1996): 557-580.

Y, Baveye, Chamaret C, Dellandréa E, and Chen L. "A protocol for cross-validating large crowdsourced data: The case of the LIRIS-ACCEDE affective video dataset." *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia.* 2014. 3-8.

Ye, Genzhi, Yebin Liu, Nils Hasler, Xiangyang Ji, Qionghai Dai, and Christian Theobalt. "Performance capture of interacting characters with handheld kinects." In *Computer Vision–ECCV*, 828-841. Springer Berlin Heidelberg, 2012.

Z., Zeng, Pantic M., Roisman G., and Huang T.S. “A survey of affect recognition methods: Audio, visual, and spontaneous expressions.” *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 31, no. 1 (January 2009): 39-58.

Appendix A: Interview questions

Background

Gender:
Country of origin:
Experience/ professional/ academic background
Working hours (how much time they spend with the resident)
Task of the care takers/nurses/doctors:

Explorative questions

- How can you tell a person's mood?
- How is telling an elder person's mood different from telling other people's mood?
- When you see or talk to a resident, what about his/her manner (gestures/posture/ facial expression/ level of activity/ gaze) could give you reason to worry?

Mood assessment

- If the resident is in an unhappy/ depressed/ sad mood, how can you tell?
    - Physically, what does he/she do?
- If the resident is feeling anxious, how can you tell?
    - Physically, what does he/she do?
- When you want to know how the resident is doing, how do you find out? (Ask questions, physique, etc.)
- Can you show how a person looks like when he/she is sad?
- Can you show how a person looks like when he/she is anxious/stressed?
- What kind of gestures or postures do you associate with sadness?
- What kind of gestures or postures do you associate with stress?
- Apart from gestures/posture, how does depression or anxiety affect a resident's behaviour in terms of actions/ level of activity?
- Does a resident make certain sounds when in a negative mood? (Such as sighing)

Mood assessment over time

- When you initially meet a resident, how do you assess the mood (without prior knowledge about the resident, health record, etc.)?
- Do you learn more over time?
- Does learning the personality of a person help?
- Do you assess more based on daily interaction?

Context

- What does a resident's average day look like?
- At what times of day do you regularly see the resident?
- How does a conversation/the interaction usually go between you and the resident? (What sort of things do you ask, what do they say, etc.)

- What might give a resident cause for stress?
- What might give a resident cause for sadness?

Activities

- What are the typical activities of a resident person in his/her room?
- How many hours does a person spend in the room and how much in the social/common room? (Does it vary a lot from person to person?)
- Which time of the day are the resident mainly in their rooms?
- When would you worry? (e.g. too much time in the room, no socializing)
- How does depression affect their daily activities?
- How does anxiety affect their daily activities?

Ambience

- Is the residents' mood influenced by the *weather (sunshine, rain, cold, dark)*?
- Is the residents' mood influenced by *the time of the day* (morning/evening)?
- What type of *events* can influence negatively an interns' mood? (No visitors, loneliness, death of a beloved person, phone call, bad food, sickness, etc.)

Appendix B: Scenarios

Scenario Anxiety 1

Situation

The character reads a newspaper. The newspaper contains a troubling article: a woman is missing. The character believes he knows the woman.

Instructions

The actor is sitting in his chair. After a while, he picks up the newspaper and starts reading it. When he sees the title of the article about the woman, he reads the article, getting increasingly worried. After reading the article, he puts down the newspaper. He may stand up, walk around, or sit as he pleases.

Intended moods

We expect a transition from neutral to anxious. The character is disturbed by the news, and feels like getting up and looking for her.

Scenario Anxiety 2

Situation

The character is sitting in a chair. It is lunchtime, but the caretaker who always gets him lunch is late. The character has beginning dementia, and realizes he may be mistaken about the time the caretaker should be there.

Instructions

The actor is sitting in the chair. He sometimes makes an effort to stand up but doesn't. His eyes dart around the room, looking for something. After a while, the caretaker enters with the food. The actor watches as the caretaker puts the food in front of him. Once the caretaker has left, the actor is free to choose whether to eat.

Intended moods

We expect an anxious mood. The character is wondering when the caretaker will be there, and what might be for lunch. He is also frustrated, because he cannot get lunch on his own and because he realizes his dementia is causing him to lose his grip on his life. We expect that, once the food has arrived, the person will calm down, ending in a neutral mood.

Scenario Depression 1

Situation

The character is in his room. Each week, his son pays him a visit. He will be coming this afternoon. The character gets a phone call from his son. At the end of the chat, the son tells him he won't be able to visit this week.

Instructions

The actor is sitting in his chair. The phone is on a desk. The actor picks up the phone and mainly listens. He plays out his son's dialogue in his head, and he himself has four lines of dialogue: "Hello. Yes. Oh. Yes." This conversation may take as long as the actor wishes. After the phone call ends, the actor may stand up, walk around, or sit as he pleases.

Intended moods

We expect a transition from happy to depressed. The character is happy at first, looking forward to his son's dropping by. When his son calls, he is happy to speak to him again. The news at the end of the phone call makes him sad.

Scenario Depression 2

Situation

The character is sitting in a chair. She's tired, and hasn't slept well. She's been sitting in her room the whole day. It is lunchtime, but the caretaker who always brings lunch is late.

Instructions

The actor is sitting huddled up in his chair. He mostly keeps glancing at the door. After a while, the caretaker enters with the food. The actor watches as the caretaker puts the food in front of him. Once the caretaker has left, the actor is free to choose whether to eat.

Intended moods

We expect a depressed mood. The caretaker feels lonely at the fact that even the caretaker appears to have forgotten her. We expect that, once the food has arrived, the person will return to a neutral mood.