Multimodal affective human-machine interaction

Towards the design and implementation of an affect-sensitive empathetic agent

- Master Thesis -

Master Artificial Intelligence

Tessa Verhoef
1404148
tverhoef@ai.rug.nl

December 31, 2008

Internal advisors:
dr. F. Cnossen. Artificial Intelligence, University of Groningen
drs. T. van der Zant. Artificial Intelligence, University of Groningen

External advisor:
dr. C. L. Lisetti. School of Computing and Information Sciences, Florida International University

university of groningen

Abstract

Present research in Artificial Intelligence brings us ever closer to a reality in which machines and robots assist humans in their everyday lives. These machines will be tutoring us, be concerned with our health, taking care of our elderly and play the part of a companion in our homes. This development increases the need for good human-machine interaction. The field of *affective computing* addresses the importance of emotions in human-machine interaction. This thesis aims to provide an overview of this field and to provide possible solutions for the design and implementation of an affect-sensitive empathetic agent that can sense and interpret the affective state of a user (affect recognition) and reacts appropriately by showing an empathetic response (affect generation). A data set was created by exposing participants to emotion eliciting stimuli while their physiological signals and facial expressions were recorded. This data was used to train and test several emotion recognizers, using both discrete and continuous emotion representations. Continuous classification appeared to be more successful than the use of discrete categories. In addition, an anthropomorphic avatar was implemented to express twelve different psychologically grounded facial expressions. These expressions were tested for recognizability and believability in a small user study. The success rate differed for each expression, but all twelve were recognized with a percentage above chance level.

Contents

"*The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without emotions.*"

- Marvin Minsky, 1985 -

Chapter 1

Introduction

Present research in Artificial Intelligence brings us ever closer to a reality in which machines and robots assist humans in their everyday lives. These machines will be tutoring us, be concerned with our health, taking care of our elderly and play the part of a companion in our homes. This development increases the need for good human-machine interaction. Obviously we should use human-human communication as an inspiration for designing machines that interact with humans in order to make the interaction both natural and effective.

In humans, emotions provide an essential ingredient to enable thought processes that influence social interactions (Nasoz et al., 2004). The role of emotions has been neglected in the design of autonomous machines for quite a long time, which does not do justice to this phenomenon, given its important role in human functioning. For a long time, researchers thought that emotions and rational thinking were strict opposites and that emotions were influencing our rationality in a negative way. Emotions were considered to get in the way of our rational thinking. More recently it has been shown that the opposite is true. Emotions are even needed for rational thinking and people with an impaired emotional system have difficulties making rational decisions (Damasio et al., 1994; Koenigs and Tranel, 2007). Other cognitive processes are also influenced by emotions, such as memory and learning. People perform better at recalling things they learned when they are in the same mood as they were while learning it (Bower, 1981). Perception also depends on our affective state. When we are sad or angry, it is more likely that we focus our perception on the negative events in our surroundings, whereas being happy makes us notice the positive things more (Bower, 1981).

In human to human communication, the role of emotions is essential. The messages people want to convey to each other are often not only reflected in the exact words they use, but the real meaning of the message is conveyed trough more subtle, non-verbal channels. Mehrabian (2007) claimed that the true meaning of a message is expressed in language for only 7 percent, while the roles of facial expressions and vocal intonation are 55 and 38 percent respectively. Although it is hard to ascertain that these numbers represent the truth, the idea is supported by the big popularity of *emoticons* in online chatting and e-mailing as well as the fact that human-human communication can be challenging when the use of social and emotional cues is impaired, for instance in the case of autism. All this indicates that it might be important to use human

emotions as an inspiration when designing machines that have to act in complex environments where natural human-machine interaction is essential.

Humans should be able to communicate with machines in the same way as they are used to communicate with other humans, so that there is no need for the user to adjust to the machine, but the machine will adjust to the human user. In fact, people already perceive machines and robots as being social actors and having a personality (Dryer, 1999). Picard (1999) states how people do not only feel emotions, such as frustration, with machines, but they also show it. It does not surprise us to find people yelling at their computer or expressing curse words toward it. 'People naturally express emotion to machines, but machines do not naturally recognize it' (Picard, 1999). Experiments have been conducted to investigate the influence of the type of interaction on the successfulness of the interaction. Walker et al. (1994) for instance compared the behavior of people answering questions via text displays and people doing the same via interaction with a talking face and they found that people who interacted with the face made fewer mistakes, wrote more comments and invested more time on the questionnaire than those who used the text display.

Branched from Artificial Intelligence in the 1990's, a new research field, called *affective computing*, emerged that addresses the importance of emotions in human-machine interaction. Research in affective computing facilitates the design and construction of machines or systems that can recognize, process and simulate emotions. It is 'computing that relates to, arises from, or influences emotions' (Picard, 1997). Affective interfaces will be applied in varying areas. In e-learning, a tutoring agent that can recognize when a student is bored, frustrated or pleased can adapt its response and make the learning process more pleasant and effective for the student (Elliott et al., 1997; Nasoz and Lisetti, 2006). Emotions play an important role in motivation, which is very important in learning. Tele-home health care also benefits from affective cues. Health care and counseling improves when the provider is aware of the user's emotional state (Rosis et al., 2006; Nasoz and Lisetti, 2006). It provides an indication of the patients mental status which can be used to adapt to the user and be more supportive. The field of robotics can also benefit greatly from research on affect. Until very recently, research in the field of autonomous perceptive systems was very much directed towards cognitive skills such as planning, searching and problem solving (Russell and Norvig, 1995). In some applications, such as surveillance or space robotics, these skills might be enough, but when robots are going to be used in elderly care, rehabilitation, therapy or as a companion or service assistant in our homes, they become an important part of our lives. In these cases the robot needs to be social and able to interact with humans in a natural and pleasant way. Especially in the field of elderly care, where due to *population ageing* care givers have less and less time available for every individual, robots will not only aid in the nursing tasks, but they will also partly substitute the social interaction that elderly people used to have with their care givers.

This thesis aims to provide an overview of the field of affective computing and to provide possible solutions for the design and implementation of an affect-sensitive empathetic agent that can sense and interpret the affective state of a user (affect recognition) and reacts appropriately by showing an empathetic response (affect generation).

The next chapter provides a background on the development of affect-sensitive

human-machine interaction systems. Several design considerations and methods that have been proposed in the past will be addressed. Chapter 3 introduces the present research which will be discussed in more detail in the remaining chapters of the thesis. Chapter 4 provides a detailed description of the first of two important building blocks in the design of affect-sensitive human-machine interaction systems: Affect recognition. This chapter shows how a data set of multimodal affect measurements was created and used for the training and testing of several affect classifiers. Chapter 5 describes the second building block: Affect generation. This chapter handles the introduction and testing of an avatar which shows psychologically grounded facial expressions. The thesis will end with an extensive overall discussion and conclusion.

Chapter 2

Background

The development of an affect-sensitive interaction system is a highly interdisciplinary task. It not only involves the implementation of an automatic emotion recognizer, but it also requires careful consideration of theories about emotions and research in psychology. Before we can model the emotions of a user, we need to specify exactly how we define 'emotion', therefore we need to take emotion theories into account, which is challenging since theorists in this field have not reached a consensus about the nature of emotions yet. In order to define what kind of evidence the system can use to make predictions about the user's affective state, empirical data of experiments in psychophysiology need to be reviewed so that modalities can be identified that give useful information. Before data can be recorded from human users, stimuli need to be created that elicit specific emotions in order to be able to label affective states correctly. Finally, it also needs to be determined how the information from multiple modalities can be combined into one decision, which means that a suitable fusion method has to be used. First, these important design considerations are discussed and then a brief overview is provided of systems that have been created in the past, focussing on the ones that incorporate multiple modalities.

2.1 Design considerations

2.1.1 Emotion theories

Reviewing all the work that has been conducted in the study of human emotions, we can go back as far as to the ancient Greeks, when Aristotle described 'akrasia' (incontinence) and the influence of 'pathos' (emotion, feeling) which was considered a direct opposite of reason. An akratic person would not act in accordance with reason, but would act under the influence of a passion (Kraut, 2007). Many great philosophers and psychologists have touched the subject of emotions since then. Most of the theories and emotion models that resulted from their work have been designed to study emotions in general. For the purpose of affective computing and the ability to automatically recognize human emotions, we need to find a structured representation of emotions that would enable us to translate the emotion theory into an engineering framework.

Discrete versus Continuous In contemporary theories that define the structure of emotions, there are several characteristics that can be used to distinguish the different theories from one another. One of these is the distinction between theories that describe a discrete set of basic universal emotion categories (Ekman, 1992; Panksepp, 1982; Izard, 1971; Tomkins, 1962) and theories that assume continuous representations of emotions, for instance in a multi-dimensional affect space (Russell, 1980; Russell and Mehrabian, 1977; Cowie et al., 2000; Watson and Tellegen, 1985). By studying the facial expression of emotions across different cultures in the world, Ekman (1992) found that there are certain universal properties connected to human affect. Travelling around the world with a selection of pictures showing a range of facial expressions, Ekman (1992) found that there are at least six basic emotions that are recognized by people all the way from Brazil to Japan as well as in the most deserted villages in the jungle of New Guinea. This idea led many researchers to try to identify a universal set of basic emotion categories. The six emotions of Ekman (1992) for instance, are anger, disgust, fear, joy, sadness and surprise. Other researchers have identified different sets which contain either less categories, such as Panksepp (1982) who names only four basic emotions, or more categories, like for instance Izard (1971) (10 categories) or Tomkins (1962) (9 categories). The fact that researchers who believe in a discrete set of basic emotions can not agree on exactly which ones or how many this should be is sometimes used as evidence against discrete categorization, but on the other hand, these researchers want to draw attention to the fact that many emotions occur in almost every list (Ortony and Turner, 1990), which shows that there is also a great amount of agreement between the different sets. There are researchers who do not believe that emotions can be grouped into discrete categories. They try to find evidence against this view and introduce alternative theories. Wierzbicka (1992) for instance, found that the words we use to indicate emotion categories can not be directly translated into other languages. There are languages that do not have a word for the state which we call 'anger', and they have words for states that we could not describe with one word. They propose the use of a meta-language to describe emotional states. Researchers who prefer to represent emotions as continuous instead of discrete categories have collected empirical evidence that contradicts the strict categorization. Russell and Barrett (1999) and Russell (1980) describe how membership in each emotion category is a matter of degree rather than all or none. By analyzing human self-report and laymen conceptualizations of affect they found that the boundaries of affect words are fuzzy. They propose a non-discrete representation in which the place in a multi-dimensional affect space defines the emotion. The dimensions are valence (pleasant/unpleasant) and arousal (active/passive) in the two dimensional case (Russell, 1980) and in the three-dimensional case the dominance (potent/submissive) dimension is added (Russell and Mehrabian, 1977). An example of such a two dimensional structure is shown in figure 2.1. Similar structures are the activation-evaluation space, as described by Cowie et al. (2000), and the two-factor structure of affect of Watson and Tellegen (1985). Lang (1995) describes affects as being organized around a motivational base in which two motive systems exist, appetitive and aversive. Combined these can be compared to the valence dimension, while arousal is assumed to reflect the variations in activation in these systems.

Challenging the categorical approach, Ortony and Turner (1990) describe how what we call, for instance, 'fear' can refer to very different affect states.

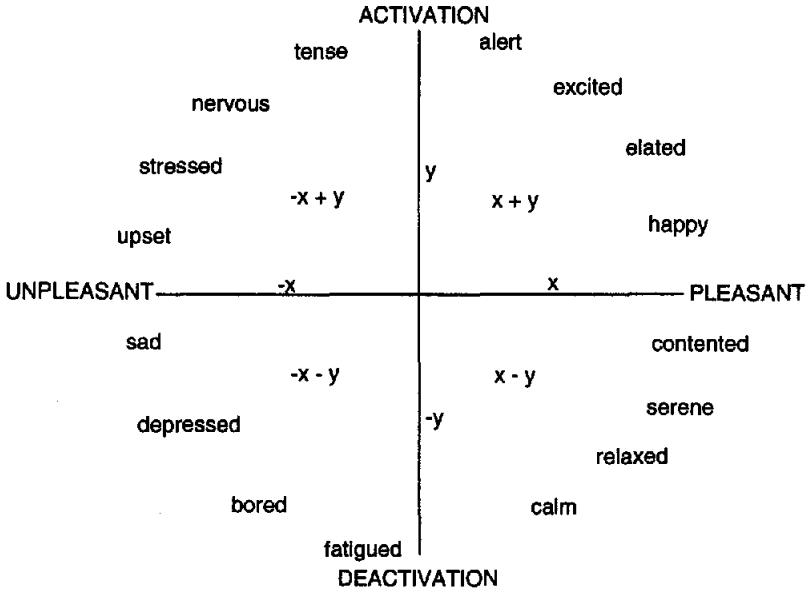


Figure 2.1: *Two-dimensional representation of affect. From (Barrett and Russell, 1998)*

When you are chased by a dangerous animal it creates a different type of fear from the case in which you would fear to have cancer. These varieties of fear also include different facial expressions and bodily responses. Therefore they say we can not use the emotion words as basic emotion concepts. Instead they describe another alternative method in the form of a model (Ortony et al., 1988) in which lower level dissociable components: specific appraisals and their corresponding responses, are considered. In this model emotions consist of valanced affective reactions which are evaluations of events, actions and/or objects. These evaluations can be compared to the kind of evaluations that need to be done to define the place in the dimensional models that were previously mentioned. In this view, appraisal theories can be considered as multidimensional with more dimensions than two or three. Scherer (2001) describes another theory based on appraisal in which he defines the *component process model*. Emotions in this model are defined as "an episode of interrelated, synchronized changes [...] in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism" (Scherer, 2001). A collection of *sequential evaluation checks* is described that drive the appraisal of situations/events and helps the organism to prepare an appropriate reaction. The 16 different evaluation checks together create a 16-dimensional affect space and the theory describes how the appraisal of a stimulus influences our physiological features such as facial expressions, vocal intonation and the *autonomic nervous system* (Scherer and Ellgring, 2007).

Affect before or after cognition? Another point that is often present in the discussion about emotions can be compared to a chicken and egg problem:

does emotion precede cognition, or does cognition influence affective states? For a very long time, it was assumed in psychology that cognition and reasoning are needed before an affective state occurs. Before we can like, love, hate or detest something, we need to know what it is and think about it, they would argue (Zajonc, 1980). Ortony et al. (1988) deny the possibility of unconscious emotions on a functional basis which views emotions as signals to focus attention on important issues. They claim that if an unconscious emotion would appear and conscious awareness of this feeling would be absent, it would be impossible to initiate any needed action. Conscious awareness should give the organism a chance to respond in an appropriate way. Lazarus (1984) adopts a similar view:

> "*When we cognize an event as pleasant or unpleasant, we are not experiencing an emotion. However, when we further cognize that we are or may be personally benefitted or harmed, the cognitive transformation has gone beyond the mere registration of discomfort, and the experience becomes an emotion.*" Lazarus (1984)

In contrast to this view, Zajonc (1980) studied the aspects of affect and feeling that are involved in preferences and argued that cognition and affect are two separate systems and that feelings come first. Affective reactions are said to be instantaneous and automatic and the empirical evidence for this view is provided in the form of a judgement task. People were asked to give a judgement of pleasantness about both faces (emotionally charged) and non-emotionally charged objects and they found that the emotionally charged judgements were accompanied by deeper information processing and resulted in better performance on a recognition task with the same faces and objects (Zajonc, 1980). Bechara et al. (1997) show experimentally how people sometimes know what is right before consciously knowing it. They investigated the behavior of both normal subjects and brain-damaged subjects (with decision-making defects) in a gambling task and found that normal people adopted the advantageous strategy before they could explain which strategy worked best, while brain-damaged subjects never used the advantageous strategy even though they could consciously explain which one was the best.

Both views make assumptions that sound correct, which makes it reasonable to assume that both orderings can actually happen. Leventhal and Scherer (1987) do not deny the existence of either of the two processing orderings. They assume in their *component process theory* (Scherer, 2001; Leventhal and Scherer, 1987) that there exist three levels of emotion processing: 1. Sensory-motor, 2. Schematic, and 3. Conceptual. At the sensory motor level, the primary emotional response capabilities are reflected, where emotions are processed without volitional effort, in an automatic fashion. At the schematic level, processing is involved with learned prototypes of emotional situations, that have been developed during emotional encounters with the environment. The conceptual level involves processing using memory and the capacity to "reflect upon, abstract, and draw conclusions about the environment and the emotional response to it" (Leventhal and Scherer, 1987). They assume that emotion and cognition are connected closely and that humans would very seldom experience one without involvement of the other.

Physiological or cognitive? A third characteristic that distinguishes emotion theories from one another is the extent to which they are physiology or

cognition oriented. Some theories, like the appraisal model described by Ortony et al. (1988), assume that all emotional processing happens in the head and they do not consider the effects or influences of bodily expressions. Other theories pay a lot more attention to the physiological responses that accompany affective changes. It has been shown that emotional processing does not only influence how physiological features respond but that, the other way around, the so called embodiment of emotions also influences how emotions are being processed (Niedenthal, 2007). In an experiment, facial expressions of participants were manipulated to test the influence on emotional processing. Test subjects were asked to hold a pencil in their mouth, either between their teeth (forcing a smile) or between their lips (suppressing a smile). All participants watched the same cartoon and they found that the people with the forced smile evaluated the cartoon as funnier than the people whose smiles were suppressed. Another experiment asked some participants to sit up strait with the shoulders back and others to sit in a slumped posture. All participants received the news that they had passed a test that they had completed earlier. The participants who were sitting in the slumped posture felt less proud and were in a worse mood than the others (Niedenthal, 2007). The component process theory of emotions (Scherer, 2001) is a theory that does take emotion physiology into account. It describes in detail how physiological features such as facial expressions, vocal intonation and the *autonomic nervous system* relate to the sequential evaluation checks. This theory might be more suitable for the field of affective computing because physiological changes can be recorded, for instance by a camera (facial expressions), a microphone (vocal intonation) or biosensors (heart rate, galvanic skin response, blood volume pressure) which makes it more suitable for computational purposes. After all, we need the possibility to translate the emotion theory into engineering frameworks for the purpose of automatic emotion recognition and generation.

2.1.2 Emotion elicitation

When researchers first started to work on affect recognition, there were no data sets available of signals or recordings of people having affective experiences. Very often, they asked actors to perform certain emotions to create their data sets. The actors would speak with a requested emotional intonation and they would act the requested facial expression. An advantage of this way of eliciting emotions is that you know exactly which emotion is supposed to be performed, which makes it easy to label the recorded data correctly. A disadvantage is the fact that posed emotions are often exaggerated and do not really reflect the affective variation that occurs in real life. It does not account for the fine grained details that are present in spontaneous emotional expressions. Moreover, it has been shown that the nature of posed expressions differs from spontaneous emotions. Cohn and Schmidt (2003) for instance found that timing differs between posed and spontaneous smiles and Valstar et al. (2006) showed that the same counts for brow actions. They could distinguish between these two conditions using an automatic classifier with high recognition rates. An alternative to asking actors to pose expressions is to elicit spontaneous emotions by exposing people to emotion charged stimuli. This should result in genuine emotion expressions that reflect real life, but it is harder to validate that the intended emotion was really elicited, which makes it more difficult to label the data correctly. A common

solution would be to ask people to self-report on the experienced emotion. This is being used a lot, but there are researchers who doubt the reliability of this method. The stimuli should be selected very carefully and preferably be tested beforehand to be sure they elicit the intended emotions. The stimuli should also be neutral to desensitization throughout the experiment, since it has been shown that affective reactions diminish after repeated exposure to similar stimuli, which is used to treat phobia's (Lang et al., 1970). Examples of emotion eliciting stimuli that can be used are movies (Nasoz et al., 2004), music (Wagner et al., 2005), interview settings (Zeng et al., 2006) or games/puzzles (Kapoor and Picard, 2005; Merkx et al., 2007; Conati, 2002).

2.1.3 Gathering evidence about affective states

To be able to recognize the affective state of a user, we need to define which modalities provide us with evidence and cues about the emotions of the person. At first, the focus was mainly directed to the use of videotaped facial expressions and audio recordings of vocal intonation (Pantic and Rothkrantz, 2003). Soon, other modalities such as physiological signals were also considered and currently researchers are still trying to discover new modalities that provide evidence about the affective state of a person.

Facial expressions The face plays an extremely important role in human functioning. Identification of the people around you and human to human communication are both skills in which the face is almost indispensable. It has been suggested that during communication between humans, 55 percent of the information is transferred by facial expressions (Mehrabian, 2007). Therefore the face provides an important source of information for emotion recognition.

Many methods have been proposed for the automatic recognition of facial expressions. They can roughly be classified into two categories: holistic methods and feature based methods. Holistic methods take the entire face into account and do not look at individual features in the face, such as the nose, mouth or facial muscle features. Feature based methods extract individual local features from the face before classification. An example of the kind of features that can be used is the Facial Action Coding System (FACS) (Ekman and Friesen, 1978). This system describes facial expressions by Action Units. Most of these action units are anatomically related to the contractions of specific facial muscles. Figure 2.2 shows some examples of Action Units and the corresponding facial muscle changes around the eyes. Local methods have the advantage that they can take a lot of detail into account in a small area of the face and are therefore able to detect fine grained changes in facial expressions. Nonetheless, many researchers prefer the holistic approach because local methods have the disadvantage of being bad at generalization. It is much harder to distinguish between changes caused by differences in lighting and changes caused by differences in facial expressions when the global overview is missing (van Kuilenburg et al., 2005). Another distinction that can be made is that of model based methods and image based methods. The model based approach creates a model of either the entire face (in the holistic case) or individual facial features such as the mouth or the eyebrows (in the feature based approach). The image based approach uses (a compressed version of) the original pixel information in the image.

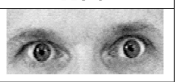
NEUTRAL	AU 1	AU 2	AU 4	AU 5
				
Eyes, brow, and cheek are relaxed.	Inner portion of the brows is raised.	Outer portion of the brows is raised.	Brows lowered and drawn together	Upper eyelids are raised.
AU 6	AU 7	AU 1+2	AU 1+4	AU 4+5
				
Cheeks are raised.	Lower eyelids are raised.	Inner and outer portions of the brows are raised.	Medial portion of the brows is raised and pulled together.	Brows lowered and drawn together and upper eyelids are raised.
AU 1+2+4	AU 1+2+5	AU 1+6	AU 6+7	AU 1+2+5+6+7
				
Brows are pulled together and upward.	Brows and upper eyelids are raised.	Inner portion of brows and cheeks are raised.	Lower eyelids cheeks are raised.	Brows, eyelids, and cheeks are raised.

Figure 2.2: *Action Units. From (Tian et al., 2001)*

An example of a facial expression recognition system that is both model based and uses local features is that of Tian et al. (2001). In this method, features are extracted from Multistate Face Component Models (Tian et al., 2001) in which a distinction is made between permanent features, such as wrinkles that appear from aging, and transient features that appear with facial expressions. FACS Action Units are recognized using these features. An image based method that is proposed by several researchers is the use of Gabor Wavelets. Bartlett et al. (2005) for instance uses a selected subset of Gabor filters to recognize seven basic emotions in frontal faces and uses the same method to also attempt a feature based approach by recognizing Action Units. An example of a model based method is the Active Appearance Model, which combines texture information and shape information and was first described by Cootes et al. (1998). This approach is used by van Kuilenburg et al. (2005) to train a holistic basic emotion classifier, which is currently commercially available. Another example of a system that uses very local information is the frustration classifier of Kapoor et al. (2007). Here, separate classifiers are used to detect fidgets in the areas around the mouth, eyes or pupils and to detect head nods.

Vocal intonation When the affective state of a person changes during a conversation, the listener can hear this from the sound of the persons voice. Prosodic features are affected by our emotions and this provides evidence about the affective state of the people we communicate with. Some examples of features that can be used from speech are: pitch, intensity and speech rate (Pantic and Rothkrantz, 2003). The pitch is directly related to the rate at which the vocal cords vibrate and tends to increase when people are more exited, either in a negative (angry, fear) or a positive (happy) way. The intensity, or energy, of the speech does not increase in the case of fear, but it does when the person is angry

or happy. The speech rate increases with excitement and in the case of sadness the opposite happens for all of these features. When speech is used to recognize emotions, the prosodic information is sometimes combined with other features that can be derived from speech, such as spectral features represented with for instance MFCC (mel frequency cepstral coefficients) (Kwon et al., 2003), lexical cues that can be found through key-word spotting and syntactic information (out-of-grammar and incomplete sentences in certain affective states) (Zhang et al., 2004). It has been shown that pitch and intensity contribute the most to distinguishing between different emotions (Kwon et al., 2003), but this is not always the case. Most researchers use data sets in which actors have uttered the emotional speech in a laboratory setting. This often results in exaggerated intonation and it is not necessarily true that what actors produce in these controlled settings is related to real emotional speech (Batliner et al., 2003). Therefore, it is more beneficial to try to obtain spontaneous emotional speech. Batliner et al. (2003) used both controlled and spontaneous speech, compared the results, and found that the dominant role of pitch and intensity diminishes as speech becomes more spontaneous.

Physiological signals Emotional processing is involved with the *autonomic nervous system*. Two separate parts can be distinguished in the autonomic nervous system: the sympathetic nervous system and the parasympathetic nervous system. Activation of these systems influences physiological features such as the heart rate, the pupil diameter, the sweat glands and the blood vessels. The sympathetic nervous system has the function to prepare the body for taking actions in sudden and unexpected situations, such as running from a dangerous animal or dealing with violence. This is also called the 'fight or flight' response and it involves dilation of the pupil, stimulation of the sweat glands, dilation of the blood vessels in large muscles and an increase of the heart rate. The parasympathetic nervous system has the function to bring the body back from the emergency status to a resting status, 'rest and digest'. It involves pupil constriction, decrease in heart rate and stimulation of organs that are important in food digestion such as the salivary glands and stomach. The autonomic nervous system is regulated by the hypothalamus, which is partly responsible for dealing with emotions. This means that emotional processing can be measured by monitoring a persons physiological signals such as heart rate, blood volume pressure, galvanic skin response and pupil diameter.

Nasoz et al. (2004) proposed a framework for modeling user's emotions from the sensory input of temperature, heart rate and galvanic skin response and attempted to map these signals to a set of basic emotions: Sadness, Anger, Surprise, Fear, Frustration and Amusement. The recognition performance varied between the different emotions, but showed to be promising. Picard et al. (2001) used the physiological signals of facial muscle tension, blood volume pressure, skin conductance and respiration. They made measurements examining a single subject over many weeks of time and found that the features of different emotions recorded on the same day sometimes clustered more closely than features of the same emotion over different days. This day-dependence shows very well how sensitive the method is for slight variations in placement of the sensors, mood effects and influences of caffeine, sleep, hormones and such. They also show that there are ways to deal with this and that it is still possible to gain a

high recognition performance. As a last example, Kim et al. (2004) used children aged from five to eight years as subjects and measured their skin temperature variation, electrodermal activity (skin conductance) and heart rate. They used many subjects, but classified into only three and four emotion categories.

The use of physiological signals is sometimes criticized because it involves wiring the user, which can be considered invasive. However, people might consider the use of video even more invasive, since it reveals the users identity, appearance and behavior (Picard et al., 2001). Moreover, wireless and wearable devices have been developed lately that are minimally invasive. Nasoz et al. (2004) for instance, use a non-invasive wearable computer BodyMedia SenseWear armband and IBM has proposed to take the measurements from an 'emotion mouse' (Ark et al., 1999). In this case there is no need for additional intrusive measuring equipment and computer users are already used to touching the mouse.

An advantage of the use of physiological signals over for instance video or audio is that emotional states are inherently activating the autonomic nervous system and therefore also inherently reflected in the bodily responses. Facial expressions and vocal intonation on the contrary can be suppressed voluntarily relatively easily. This is something that could even happen unconsciously under the influence of the awareness of being monitored.

Other modalities Researchers are still trying to explore the use of additional modalities that can provide even more evidence about the affective state of a user. Puri et al. (2005) for instance found that there is a relation between stress and increased blood flow in the area around the forehead and they proposed to use thermal imaging to detect this change. Kapoor and Picard (2005) used a special sensing chair to keep track of the users' body posture and used this as a cue to determine their emotional state. They also included head movements as evidence. Another example is the use of context, for instance the state of the program the user is working with or the progress on a task (Conati, 2002) and also the interaction with the input devices such as mouse movements and keystrokes (Maat and Pantic, 2007).

2.1.4 Multimodal sensor fusion

When an affect recognition system uses information from multiple sensor modalities, the data from the various sensors have to be fused in order to make a decision about the affective state. Three different multimodal fusion methods can be distinguished (Pantic and Rothkrantz, 2003; Paleari and Lisetti, 2006b; Sharma et al., 1998). First, signal (or data) level fusion can be used, in which the raw input, obtained directly from the sensors is combined. The next step would then be to extract features from this data assemblage. A disadvantage of this method is that it needs the multiple sensors to create data of the same type. In stereo vision, for instance, the two camera images could be combined in this way, but in multimodal affective computing it is hardly ever applicable since the various modalities differ in nature a lot. Input from video, audio, physiological signals and such create data with different temporal structures and signal characteristics (Paleari and Lisetti, 2006b). Second, there is feature level fusion which combines the features that are first extracted from the raw data of the multiple inputs. These features are combined in a joint feature

space that can be high dimensional and can cause the classification to become computationally expensive (Sharma et al., 1998). Techniques for doing feature level fusion include Kalman Filters, Artificial Neural Networks, Hidden Markov Models and Bayesian Networks (Sharma et al., 1998; Pantic and Rothkrantz, 2003). Third, the modalities can be considered as separate classifiers that produce scores individually and these scores can then be combined at the end at the decision level. A wide variety of classification methods, including the techniques just mentioned can be used for this. Decision level fusion can deal with individual sensor failure the best and is generally less computationally complex than feature fusion (Sharma et al., 1998), but it has been argued that this fusion method is almost certainly incorrect (Sebe et al., 2006; Pantic and Rothkrantz, 2003) because humans display affective signals through multiple modalities in a complementary and redundant manner. If the goal is to achieve biologically plausible, human-like affect recognition, then feature level fusion should be used. Nonetheless, so far decision level fusion has been used the most. In a case where both decision and feature level fusion have been tested, the performance was similar for both methods (Busso et al., 2004), so it is basically undecided which method would be the best one in general and researchers should base their choice on their specific goals and needs.

2.2 Affect-sensitive system solutions proposed in the past

Now that we have discussed some of the general challenges and design considerations in affective computing, we can take a look at affect-sensitive system solutions that have been proposed in the past. The design and development of complete affective interaction systems in which the user's emotional state is recognized, processed and interpreted and an appropriate adaptive response is generated has been practiced only since the last couple of years. There are examples of such complete systems, but there is also a large amount of literature in which the focus lies only on parts of the problem, either the affect recognition part or the way in which the system should respond to the human user, for instance by emotion generation with an affective display.

2.2.1 Affect recognition

Around the time that Picard (1995) published a technical report called 'Affective Computing' in which she first introduced the idea of this research field, researchers started to investigate the possibilities of automatic affect recognition by a machine. At first the focus had been directed towards methods that use cues from only one modality. Mainly this involved either facial expression recognition or voice intonation cues. Soon, inspired by the way humans integrate information from multiple modalities, the idea of multimodal affect recognition was introduced. Apparently the use of multiple sensors also yielded better performance (Sebe et al., 2006; Pantic and Rothkrantz, 2003). Sharma et al. (1998) discuss several (biological, practical and mathematical) reasons why multimodal affect recognition should be preferred over unimodal systems. Humans as well as other organisms integrate information from multiple sensors all the time. In addition to the fact that it is biologically plausible, it also makes a system more

robust and accurate. It is now widely accepted that affect recognition should be done in a multimodal fashion, therefore we focus only on past work in which multiple modalities are integrated. Pantic and Rothkrantz (2003) provide a survey in which systems are discussed that focus on either facial expressions or voice tone.

When researchers in emotion recognition started to fuse multiple modalities, many systems became bimodal, integrating facial and vocal expressions. Sebe et al. (2006) for instance combine audio and visual signals to make a classification into 11 discrete emotion categories. The signals are fused using a Bayesian Network. Caridakis et al. (2006) also use facial and vocal cues, but they classify into a dimensional emotion representation, the 2D Activation-Evaluation space. Fusing was done with a recurrent Neural Network, which allowed the system to 'memorize' past states and take temporal data into account. Zeng et al. (2007) implemented another bimodal audio-visual affect recognizer that classifies into 11 categories and they fuse the signals using Hidden Markov models. A variety of combinations of more than two modalities have also been reported. Fragopanagos and Taylor (2005) for instance combine audio and video with lexical content in order to recognize emotions in the 2D Activation-Evaluation space. Paleari and Lisetti (2006b) propose a modular and adaptive architecture in which cues from the face, voice and physiological signals are combined and Kapoor and Picard (2005) fuse facial expressions with head gestures and body posture. Barreto et al. (2007) combine physiological signals with measures of the pupil diameter to detect stress. Due to the big differences in the nature of the experiments and the choices of emotion representation and elicitation methods, it is impossible to compare the recognition results of the above methods in a meaningful way. Therefore it is still undetermined which fusing methods or machine learning techniques are the most promising.

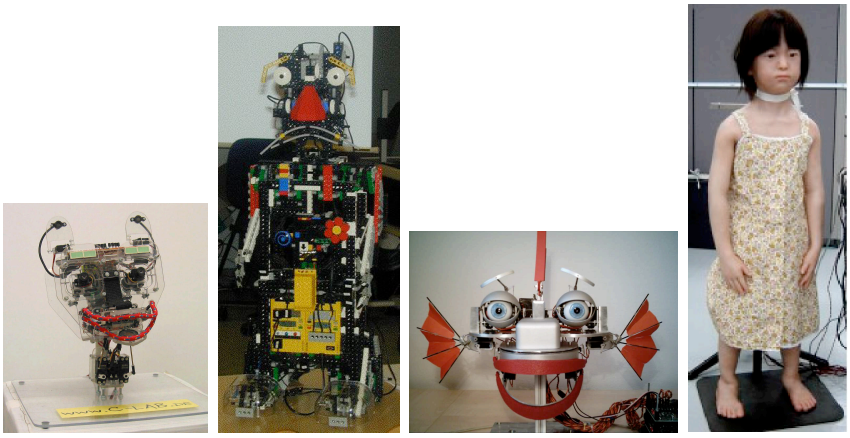
2.2.2 Affect generation

In order to realize human-like and effective human-machine interaction, the system should not only be able to recognize the affective state of the user, but it also has to show its own emotions to the user, or in our case, show its empathy. The absence of these expressions could be interpreted as distant and cold by the user (Bartneck et al., 2004), which is what we want to prevent. There are a lot of different ways in which a machine can display affective states: from the colored lights on Sony's Aibo[1] to complete affective robotic faces (Breazeal and Scassellati, 2000; Esau et al., 2003; Sosnowski et al., 2006; van Breemen, 2004; Canamero and Fredslund, 2000) and even human-like robots and avatars showing human-inspired facial expressions (Minato et al., 2004; Lisetti and Maurpang, 2006; Lisetti et al., 2004; Bruce et al., 2002).

Sony's robotic dog, Aibo, has a very simple mechanism for displaying emotions. A group of led lights that can show different colors makes the robot express its feelings. A similar method can be found in Pod[2], an emotions-expressing car with a personality. It is happy to see you like a friend when you approach it and it turns red if your driving behavior is bad. Of course lights provide a very limited way of showing affect. Affective robot faces also exist,

[1] http://support.sony-europe.com/aibo/
[2] http://www.toyota.co.uk/infront/products/pod.htm

that are able to display emotions through facial expressions. Kismet (Breazeal and Scassellati, 2000) may be the best known example of a robotic emotion display. It is able to interact with humans in a way that is based on the interaction between infants and caregivers, by watching and listening to the human companion and expressing for instance vocal babbles and facial affective cues. Another robotic display, that show some similarities to Kismet, is MEXI (Esau et al., 2003) (shown in figurefig:mexi), a robot that adopts a behavior based approach for action control and interacts with humans in a way that is recognized as human- or animal-like. Also similar is EDDIE (Sosnowski et al., 2006) (shown in figure 2.3(c)), an 'Emotion Display with Dynamic Intuitive Expressions', in which the actuators are directly assigned to particular Action Units. A commercially available example of an affective display is the iCat (van Breemen, 2004), developed by Philips. This user interface robot has the shape of a cat and is able to express emotions through both facial expressions and colored lights. It is a research platform that is aimed for developing family companions. A last example is Feelix (Canamero and Fredslund, 2000) (shown in figure 2.3(b)), an affective robot that is built from LEGO and interacts with a user through tactile stimulation. The previously discussed robotic interfaces all represent cartoon-like figures that remind us of animals or fantasy figures. In contrast to that, there are also robotic displays that aim to be as human-like and realistic as possible. In Tokyo, for instance, an android robotic receptionist has been created with the ability to express fine facial expressions (Hashimoto et al., 2004). A system with pneumatic actuators is used to resemble human muscle movement as closely as possible. In Osaka, an android with the appearance of a five year old girl has been developed (Minato et al., 2004) (shown in figure 2.3(d)). They used a mold of a real girl to make it look realistic and resemble a human girl.



(a) Mexi (b) Feelix (c) Eddie (d) Android

Figure 2.3: *Various robotic affective displays: a) Mexi, taken from (Esau et al., 2003), b) Feelix, taken from (Canamero and Fredslund, 2000), c) Eddie, taken from (Sosnowski et al., 2006) and d) Child android, taken from (Minato et al., 2004)*

Using a real embodied and situated robot can be difficult, expensive and

time consuming and the amount of degrees of freedom is always quite limited in the case of mechanical actuation. An alternative approach is to use 3D renderings of a face on a screen in which many degrees of freedom can be modeled for generating realistic facial expressions. Nasoz and Lisetti (2006) created a Multimodal Affective User Interface (MAUI) in which avatars are used to provide feedback to users about their affective states. An application domain in which affective avatars are popular is e-learning. It has been shown that the use of animated tutoring avatars can have a positive effect on the students learning experience (Lester et al., 1997). An example of such an avatar is Steve, a pedagogical agent that operates in virtual environments and assists students in learning physical, procedural tasks. Avatars have also been used in robot interfaces. Cherry (Lisetti et al., 2004) for instance is an autonomous social robot that has the ability to express her emotions through an affective avatar on an screen. Vikia (Bruce et al., 2002) is another example of a social robot that communicates with humans through a 3D rendered female face. When entering the Newell-Simon Hall at Carnegie Mellon University, the 'Roboceptionist' (Gockley et al., 2005) gives visitors information and directions about the building and the people working in it through a graphical human-like face on a screen while the user can enter questions in a keyboard. Figure 2.4 shows the Roboceptionist.

Figure 2.4: *The Roboceptionist in the Newell-Simon Hall at Carnegie Mellon University*

When creating human-like avatars or robotic interfaces we always need to be aware of 'The Uncanny Valley' as described by Mori (1970). This theory describes how the relation between the degree of human-likeness and human acceptance (or the positive emotions the creature elicits) is not an ever-increasing correlation, but is plagued by a deep valley as shown in figure 2.5. At this moment the robot or avatar is very close to human-like, but misses something which makes it strange and zombie-like and elicits revulsion instead of pleasantness.

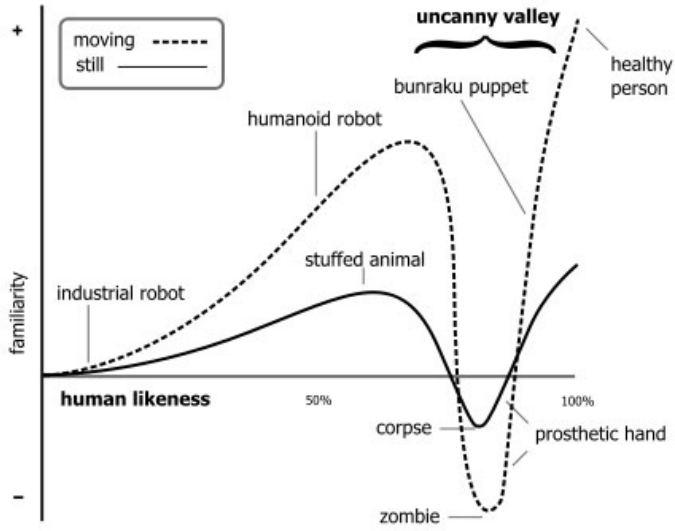


Figure 2.5: *The Uncanny Valley. From (Mori, 1970)*

2.2.3 Complete systems

One of the first complete systems that involved both emotion recognition and an appropriate adaptive response to the user's sensed affective state is described in Lisetti and Nasoz (2002). This system uses kinesthetic cues to recognize the emotion and an interface, including an avatar that mirrors the user's affective state, is adapted and provides the user with feedback. Another, more recent example is that of Melder et al. (2007), where the input of speech, facial expressions and gestures is used to detect laughter. The system reflects the image of the user in a funny way and the user can influence the mirror by laughing which induces certain visual effects. Maat and Pantic (2007) designed a system for an office scenario that used the user's facial and vocal affective cues and gaze direction as well as information about the context (task, mouse clicks, keyboard strikes) to support the user with the task. Nasoz and Lisetti (2007) proposed a system that can be used to improve driver safety. The user's affective state is recognized and an adaptive interface can respond appropriately, for instance to the driver's fatigue or anger.

A good example of a complete affective interaction system in robotics is Kismet (Breazeal and Aryananda, 2002). Breazeal and Aryananda (2002) integrated their robot with a system that could recognize affective intent in speech: approval, attention, prohibition or soothing, and the robot showed an appropriate expression in response. Their experiments showed very well how people attribute feelings to such a machine and are compassionate about a system, since some participants felt guilty for making Kismet sad. Currently the MIT media lab is working on a therapeutic robot, ShyBot (Lee et al., 2008). This robot will be used to interact with autistic children, recognizing their anxiety and making them re-experience stimuli that are stressful to them, like social interaction, to overcome excessive shyness.

Chapter 3

Present research

The research described in this thesis involves the design and implementation of the two important building blocks necessary for creating an empathetic affective agent: affect recognition and affect generation. The idea is that the affect recognition system will provide useful information about the affective state of the user that helps in the decision of how to react and adapt to the user. The affect generation system then shows the appropriate response.

Emotion recognition was done in a multimodal fashion in which affective information from different time frames was integrated: rapidly varying emotions that evolve over minutes on the one hand and longer/life-time lasting personality traits (Paleari and Lisetti, 2006b; Scherer, 2005; Kshirsagar, 2002; Pantic and Rothkrantz, 2003) on the other. It is believed that people with different personality traits also respond differently in their emotional expressions. Introverted people for instance are said to have stronger responses in their Galvanic Skin Response (GSR) and Blood Volume Pressure (BVP) signals while their facial expressions are less outgoing than those of extraverted persons (Picard, 1997). Neurotic individuals will have a higher tendency towards expressing negative emotions and aggressive persons will express emotions like anger more often than others. Knowing these typicalities about a user can provide beneficial background knowledge about how users will respond and how the measured physiology should be interpreted. Multiple sensor modalities were used to gather evidence about the affective state of the user and personality traits were obtained through a simple questionnaire. The user's physiological signals were sensed, including GSR, heart rate, BVP and facial expressions, when the users were exposed to specific emotion eliciting stimuli. In this way labeled data of physiological signals corresponding with certain emotions was obtained in collaboration with the Department of Electrical & Computer Engineering of Florida International University, where an experimental set-up for measuring physiological signals was available. The obtained data set was used to train and test several emotion classifiers.

The focus of our system adaptation is not directed towards a specific application but on the expression of empathy towards the user (affect generation), which is a first step of letting the machine show the users that it is aware of their emotions. We distinguish two different kinds of empathy: affective empathy and cognitive empathy. Affective empathy is expressed as a low-level response that can even be identified in very small children. It is an empathetic reflex that ba-

sically mirrors the affective state of the other person. Cognitive empathy refers to the higher-level responses, where people consciously understand the affective situation of the other and respond in a way that helps or supports the other person. We only focused on a simple affective empathy in our affect generation system, where an avatar was implemented to express certain emotions. Figure 3.1 shows a schematic diagram of how the two parts, affect recognition and affect generation, could be integrated to create an empathetic agent.

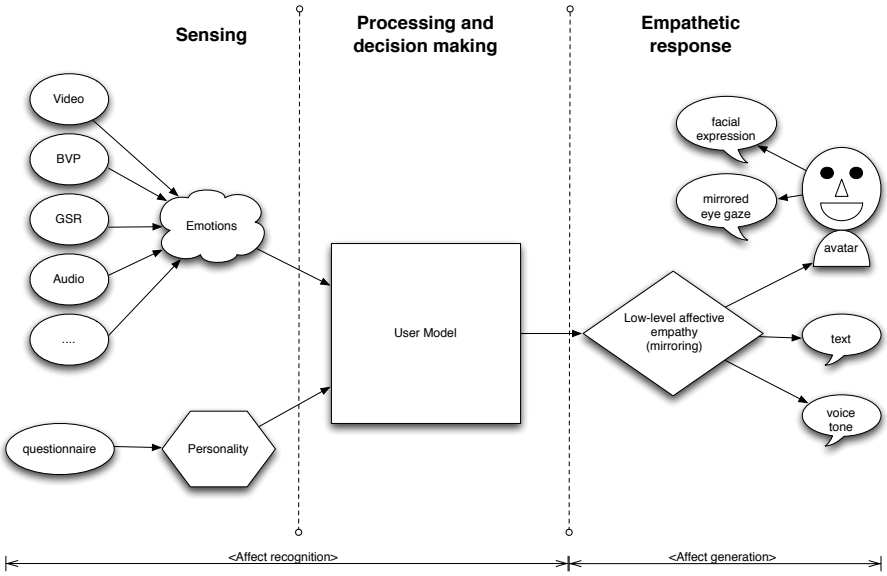


Figure 3.1: *Schematic diagram of complete empathetic agent, combining the two parts of affect recognition and generation. The affect recognition part uses both emotions and personality as input, provided by input from video (facial expressions, hand gestures, body posture), Blood Volume Pressure (BVP), Galvanic Skin Response (GSR), Audio (voice tone) for emotions and a questionnaire for personality. The affect generation part mirrors the recognized emotion through multiple modalities such as written text, a speaking voice and an avatar with facial and ocular expressions. This figure also displays that the structure can be expanded by adding more modalities both on the sensing and on the generation part.*

The two parts of the system, affect recognition and affect generation, will be described in the following chapters. The chapter on affect recognition describes how a data set of affect measurements was created which was used to train, test and compare affect recognizers based on promising machine learning techniques. The affect generation chapter describes how anthropomorphic avatars can be used to let a machine express empathy. Psychologically grounded expressions were implemented on an avatar and these were tested for believability and recognizability in a small user study.

Chapter 4

Affect recognition

In order to train an emotion classifier, we first determined from which sensor modalities to use information as evidence about the affective state of the user. Next, we created a data set suitable for training and testing emotion classifiers. The following section describes the process of data collection. The various classifiers that have been trained and tested will be discussed after that and these include Static Bayesian Networks, Dynamic Bayesian Networks and k-Nearest Neighbours. These classifiers have been used to recognize the emotions as discrete categories, following the discrete emotion theories. As an alternative approach we also followed a method that was proposed by Peter and Herbon (2006) and allowed us to use a continuous representation of affect. In this way we trained and tested another classifier. Figure 4.1 shows a diagram of the process of data collection and recognition. Each step will be explained in more detail in the following sections.

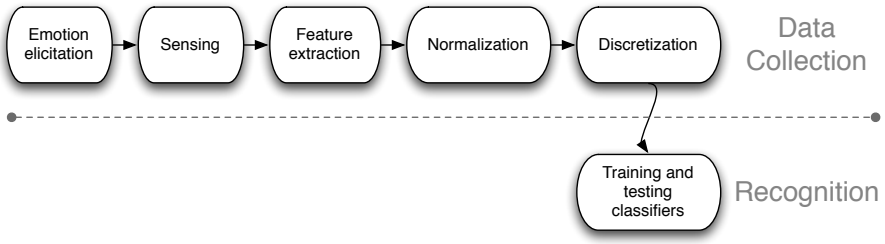


Figure 4.1: *The proces of data collection and recognition in a diagram*

4.1 Collecting data

4.1.1 Sensors

To collect data with physiological evidence about the affective state of a user we started a collaboration with the Department of Electrical & Computer Engineering at Florida International University, where Dr. Armando Barreto provided us with the possibility to make use of his biosensors. The biosensors that were

available included a GSR2 (Thought Tech LTD[1]) device and a Pulse Plethysmograph (UFI model 1020[2]). In addition to that an iBot Pro Firewire webcam (Orange Micro[3]) was used.

Galvanic Skin Response The GSR2 device (see figure 4.2) was used to measure the Galvanic Skin Response (GSR). This method has been introduced in the early 20th century and is based on the idea that conductance of an electric current is easier on moist skin. The autonomic nervous system has an influence on the control of sweat glands. In the case of higher sympathetic activity, the sweat glands get more hydrated and skin conductance increases. So, the sweat glands are used as resistors and the skin conductance can be measured with the GSR device by passing a small electric current across two elektrodes that touch the skin.

Figure 4.2: *The GSR2 skin conductance device (Thought Tech LTD)*

Blood Volume Pressure The Pulse Plethysmograph (see figure 4.3) was used for measuring the Blood Volume Pressure (BVP), a signal from which information about the Heart Rate Variability can be computed. This Heart Rate Variability has been linked to emotional processes and the autonomic nervous system (Dishman et al., 2000). In addition to this, information about vasoconstriction can be deduced by detecting a decrease in the amplitude of the BVP signal. It is a constriction of the blood vessels and is said to be related to emotional processing as well (Hilton, 1982). The device is a finger clip which uses an Infrared emitter and receiver to measure the amount of light that is reflected back by the skin.

Video of the face The iBot Pro webcam (see figure 4.4) was used to record video of the participant's face while they watched the eliciting movies. From this recording the facial expression can be deduced if the video is analysed by facial expression recognition software.

[1]http://www.thoughttechnology.com/gsr.htm
[2]http://www.ufiservingscience.com/Pig1.html
[3]Out of business since 2005

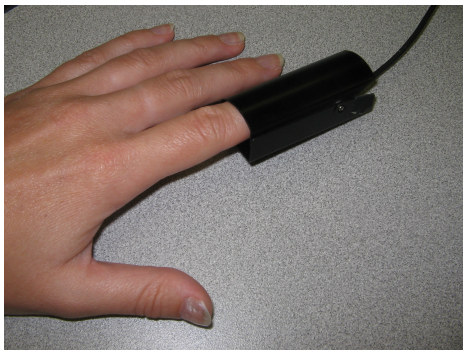


Figure 4.3: *The Pulse Plethysmograph Blood Volume Pressure device (UFI model 1020)*

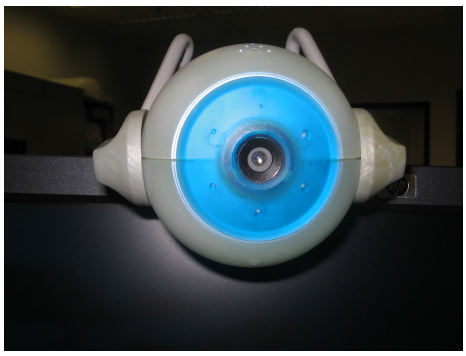


Figure 4.4: *The iBot Firewire webcam (Orange Micro)*

4.1.2 Elicitation

An experimental set-up was created with the sensors described above in which test subjects were exposed to emotion eliciting stimuli to create a data set. During the experiment, the physiological signals were measured with the non-invasive sensors while the participant was asked to keep the arm that was attached to the sensors as motionless as possible. There was a written questionnaire involved in which users were asked to fill out general demographic information and complete the Ten-Item Personality Inventory (TIPI) as described by Gosling et al. (2003).

This TIPI is a short questionnaire that includes ten questions about a persons personality traits. There are other measuring scales for personality traits available that are more precise, but these require that the user answers 40 to 100 questions. We decided to use the compact one because determining the personality was not the main focus of our experiment. Furthermore, this TIPI has been evaluated and considered adequate (Gosling et al., 2003) which makes it a suitable tool for defining the personality of our participants in the context of our experiments.

Stimulus design The stimuli that were used for emotion elicitation consisted of movie fragments that were known for eliciting a range of different emotions. Gross and Levenson (1995) conducted a very thorough study to make a selection

of movie fragments that were best suitable to elicit certain emotions. Using a large amount of test subjects and a wide selection of movie fragments they were eventually able to reduce it to a reliable set in terms of discreteness and intensity. Nasoz et al. (2004) did another panel study in which they tested this set again and this resulted in a new set which they used in their experiment which is comparable to ours. The set consisted of the following movie clips: The Champ for sadness, Schindler's List for anger, The Shining for fear, Capricorn One for surprise and Drop Dead Fred for amusement. Only Schindler's List was not originally part of the set created by Gross and Levenson (1995). In order to allow for easy comparison our set of emotion eliciting movie clips was based on the selection of Nasoz et al. (2004). Some changes have been made though because during a pre-testing stage it appeared that people responded inappropriately to some of the movies. The Shining is so well known and by now so old, that people often show a smile of recognition instead of fear. Drop Dead Fred caused people to be annoyed more than amused, probably because it goes too slow for present standards. To find out whether these two movies should be replaced a small pilot study was conducted in which we showed the two movie clips as well as two alternatives (The Ring for fear and the Pixar short movie Boundin' for happiness) and asked people to rate the emotion that they felt while watching the clip. They were asked to choose one of the following possibilities [Happy, Angry, Sad, Disgusted, Surprised, Afraid, Neutral or None of the above] and to rate the intensity of the felt emotion on a scale from 1 to 5. Fifteen test subjects participated in this pilot study among which 7 were female and 8 were male. Their ages varied from 22 to 57. The results are shown in table 4.1. The difference in eliciting success (defined by the percentage of subjects that reported to have felt the intended emotion) and average reported intensity between the two movies for 'happy' is smaller than the difference for the 'fear' movies, but in both cases the alternative movie scores better, therefore we decided to replace them both. The selection of movies for the main experiment then became: The Champ for sadness, Schindler's List for anger, The Ring for fear, Capricorn One for surprise and Boundin' for happiness and an episode of Fear Factor for disgust.

Movie clips 'happy'	Eliciting success	Average intensity	SD
Drop Dead Fred	67 %	2.8	0.79
Boundin'	73 %	3.1	1.39
Movie clips 'fear'	**Eliciting success**	**Average intensity**	**SD**
The Shining	87 %	2.6	1.14
The Ring	100 %	3.9	1.16

Table 4.1: *Results of the pilot study about eliciting abilities of movie clips. Average eliciting success, average intensity rating and Standard Deviation of the intensity rating.*

Procedure During the main experiment, the user watched the six selected movie fragments which were separated with a reasonably long pause to make sure that the subject would be totally relaxed and in a neutral state again before the next movie started. Before the movies started there was such a pause as

well, in which relaxing music was played and the participant was asked to breath slowly and try to relax. After each fragment, the user was asked to self-report on the emotion he or she felt during the movie on the written questionnaire. Two ways of self-reporting were used. The first consisted of a list of several emotions (Happy, Angry, Sad, Disgust, Surprise, Fear, Other) from which the test subjects had to choose one. The other method involved self report with the help of an 'emotion wheel', which was created by Scherer (2005) as a tool that is both intuitive and easy to use for participants. It allows to do self report with the use of discrete emotions, the way that people are used to talk about their emotions, but it also maps to a continuous 2D interpretation in terms of the two most important sequential evaluation checks as described in the component process theory of emotions (Scherer, 2001). This second way of doing self reports was used for affect recognition in the alternative, continuous approach that was proposed by Peter and Herbon (2006). Our version of this emotion wheel and the complete written questionnaire can be found in Appendix 1. The duration of the complete procedure was approximately 45 minutes. 25 test subjects participated in the experiment who varied in age from 21 to 41. The group consisted of 16 males and 9 females and the division of their ethnicities was as follows: 40 % Caucasian, 40 % Latin American and 20 % Asian. Some of the data had to be excluded from the data set. One reason to exclude data was unsuccessful elicitation. Whenever a participant self reported an emotion that did not match the intended emotion for a movie fragment, the data for that movie fragment was not used in the data set. Another reason to exclude data was unsuccessful recording of the signals. Sometimes participants moved the arm with the sensors too much which caused interruptions in the signals and this makes it impossible to compute the features in the signal. There was one special case in which the GSR signal was completely absent (too low for the range in which it was captured) except for a few seconds during one of the movie clips. The data of this participant[4] was excluded from the data set completely.

4.1.3 Feature extraction

For each test subject the experiment results in three signals: The raw GSR signal, the raw BVP signal and a quicktime movie of the recorded facial expressions. In addition to that, the questionnaire gives us the self-reported TIPI values about the personality traits. To create the data set, we compute features from the recorded signals. The features we compute from the GSR and BVP sensors are the same as the ones that are assessed by Barreto et al. (2007), with the only difference that we do not consider each movie as a complete segment from which each feature is computed over the whole segment, but we assess the signals in intervals of 40 seconds, so that we have a sequence of feature values for each elicited emotion. We had to do this because one of the classifiers that we used (Dynamic Bayesian Network) analyses temporal data for which we needed sequences of variables.

A typical GSR response consists of several temporary increases, the skin conductance responses. Figure 4.5 shows an example of such a response. Often

[4]Anecdote: the participant was a Buddhist Monk who meditates several times a day. The fact that his baseline GSR was so much lower than that of the average participant might have something to do with the control that is needed for practicing these meditations and the years of experience he has with reaching a relaxed state.

an electrodermal response is described using a few specific characteristics from these responses: amplitude, rise time and the half-recovery time (Barreto et al., 2007). The specific features that we compute from the GSR signal are the number of GSR responses, Mean value of the GSR, average Amplitude of the GSR responses, average Rising time of the GSR responses and the average Energy of the responses (the total area under the half-recovery time). All these features are computed in the way described by Barreto et al. (2007).

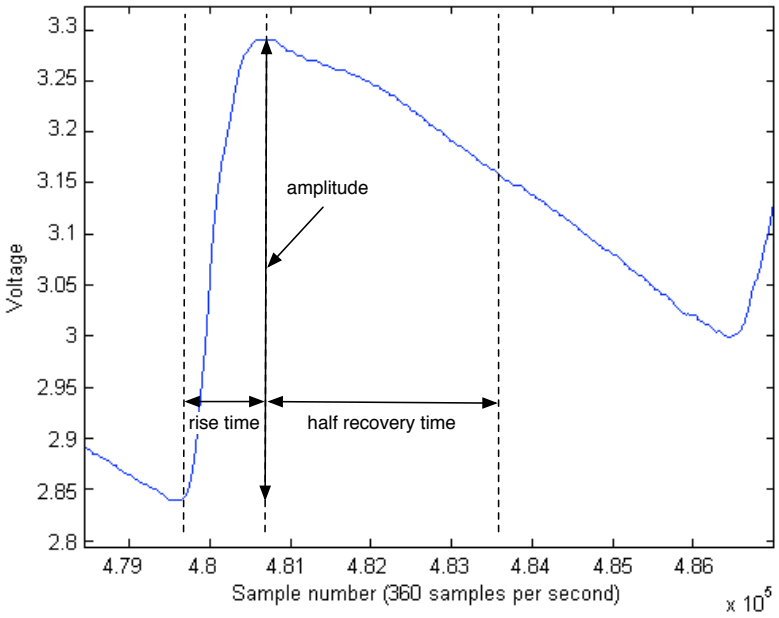


Figure 4.5: *An example of a Galvanic Skin Response and some of its characteristics*

The BVP signal can be used to compute features such as the Heart Rate Variability, like it was mentioned before. A typical BVP beat is shown in figure 4.6. As described in Barreto et al. (2007), the individual heart beats are first separated by finding the Inter Beat Intervals (or period) which is the time between two peaks in the signal. This series is usually analyzed by assessing different levels of frequency bands in which the Low Frequency (LF) (0.05-0.15Hz) band reflects sympathetic and the High Frequency (HF) (0.16-0.40Hz) band parasympathetic activity. The LF/HF ratio is computed as one feature as well as the mean Inter Beat Interval, the standard deviation of the Inter Beat Interval and the mean amplitude of the individual beats that are detected in the segment.

From the ten items in the TIPI personality test, two items relate to each of the 'Big Five' traits from the Five Factor theory of personality (Digman, 1990). Evaluating the items gives a score for every trait (Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness). The scoring that results from this is simplified to a true/false score for every trait. Among the ten items are both a positive and a negative item for every trait. The score is computed by

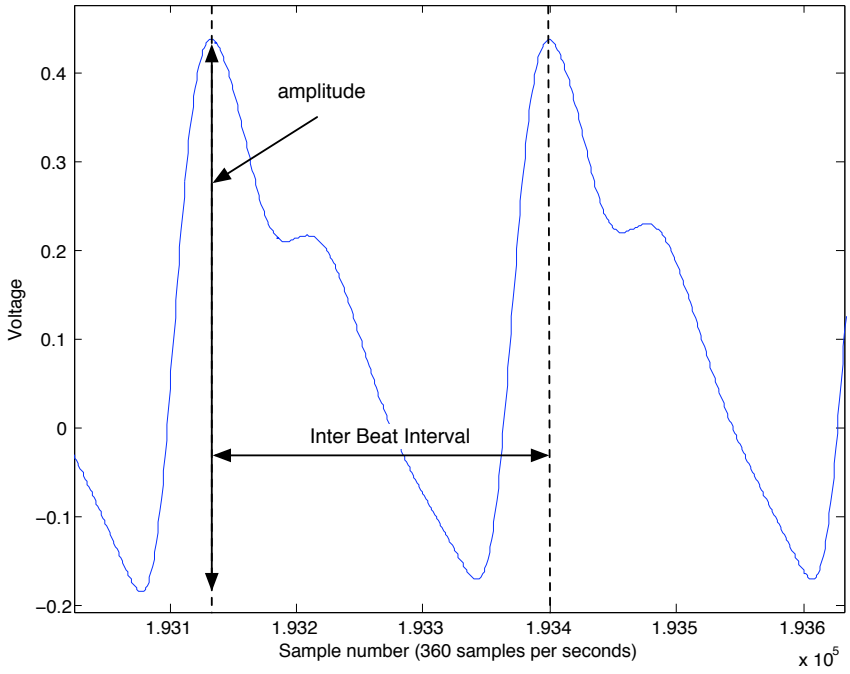


Figure 4.6: *An example of a Blood Volume Pressure beat and some of its characteristics*

taking the average of the positive and the inverse score of the negative item. In the simplification step we make use of the normative data provided by the inventors of the TIPI test (Gosling et al., 2003). The mean values of the whole sample of all ethnicities are used as a threshold to decide either true or false for each trait.

The video data of the faces is not analysed for features but is directly fed into a commercially available Facial Expression recognition system: The FaceReader (Noldus[5]). The recognized emotion that resulted from this was fused at the decision level with the result of the recognizer that fuses the other features.

Some problems were encountered with one of the six emotion categories because of the nature of the elicitation method. The duration of the surprise part in the movie clip that elicited this emotion only lasted a few seconds. The way we compute the features from the signals requires the assessment of segments of at least 40 seconds because for instance a single GSR response can last that long. Moreover, to train the Dynamic Bayesian Network, we need sequences of these segments of at least 3 slices long. This makes it unfeasible to use the signals that we recorded for surprise in the way the data is processed, so we decided to use only the data for the other 5 emotions and built a recognizer for this 5 class problem.

[5]http://www.noldus.com/site/doc200705001

4.1.4 Normalization

The physiological response to emotion eliciting stimuli differs a lot from person to person. Therefore the data from the two biosensors is normalized so that it reflects the proportional difference in reaction for the different stimuli segments and also to re-scale the individual baselines. These normalization steps follow the example of those used by Barreto et al. (2007) with similar data.

The first normalization step uses a set of features that were computed during the relaxation interval that preceded the first movie clip of the experiment. This interval represents the baseline of the user's reaction. If X_e is one of the features for the segment eliciting emotion e, and X_r is the feature recorded during the relaxation interval, then equation 4.1 computes the feature value of X_e after the first normalization step by dividing the original value by the one in relaxation.

$$X_e^{'} = \frac{X_e}{X_r} \tag{4.1}$$

The second normalization step also reduces the influence of individual differences and makes sure the baselines and strength of the responses of the individuals are equalized. If $X_e^{'}$ is one feature value after the first normalization step, for the segment that elicited emotion e, the second normalization step follows equation 4.2. The value is divided by the average individual response for that feature computed over all feature vectors n for all segments corresponding to the six emotions.

$$X_e^{''} = \frac{X_e^{'}}{\frac{1}{n}\sum_{i=1}^{n} X_{ei}^{'}} \tag{4.2}$$

The last step that is computed normalizes the features to a uniform range in order to eliminate differences in dynamic range and the chance that this makes some features dominate others. This min-max normalization step follows equation 4.3 and maps all computed feature values to a value in the range from zero to one.

$$X_{norm} = \frac{X_e^{''} - X_{e\min}^{''}}{X_{e\max}^{''} - X_{e\min}^{''}} \tag{4.3}$$

4.1.5 Discrete features

As a last preprocessing step before the data set could be used for classifier training and testing, the continuous attributes were made discrete. When a data set has continuously valued attributes the chance of overfitting is higher than when they are discrete (Holte, 1993). Also, Bayesian methods have mainly proven to be successful on discrete data (Duda et al., 1973). We use a very simple method for making the attributes discrete which is also the first step of the *1R* classifier as described by Holte (1993). When a range of values has to be transformed into a set of disjoint intervals there is a risk of using rules that are too specific for the data set. This will create too many small intervals. To prevent this overfitting, Holte (1993) proposes this simple rule: make sure that every interval (except the last one) contains more than a predefined number, the threshold, of examples in the same class. We used this rule and based on a small informal experiment we set the threshold at ten.

4.2 Classifier training and testing

A wide variety of classification techniques have been used in the past for multimodal emotion recognition. Among these are for instance Neural Networks (Fragopanagos and Taylor, 2005; Maat and Pantic, 2007; Caridakis et al., 2006) and Hidden Markov Models (Zeng et al., 2007). Barreto et al. (2007) used and compared three different machine learning techniques: Naive Bayes, Decision Trees and Support Vector Machines. Nasoz et al. (2004) used their data to classify into emotion categories with k-Nearest Neighbours, Discriminant Function Analysis and Marquardt Backpropagation. We also present several different approaches in the following sections. The performances of both Static and a Dynamic Bayesian Networks are compared on our data and this is also evaluated in comparison to a simple k-Nearest Neighbours approach. In addition to this, an alternative approach is presented in which we use a continuous emotion representation instead of a discrete categorization and train a classifier following a method described by Peter and Herbon (2006). The final data set, after the exclusion of examples, the feature extraction and preprocessing, consisted of 656 examples. 180 of these are examples of the category Sadness, 81 of Disgust, 95 of Anger, 117 of Fear and 183 of Happiness.

4.2.1 Bayesian Belief Network

Recently the affective computing community has developed an increased interest in the power of Bayesian inference. Bayesian methods have been proposed as a potential good solution more than once in the research field because it is a suitable method to deal with both noisy data (Pavlovic et al., 2000; Pantic et al., 2005) and context dependent information (Rosis et al., 2006; Pavlovic et al., 2000; Pantic et al., 2005). Emotions and behavior have such a non-deterministic relationship that uncertainty plays a large role (Ball, 2001) and Bayesian networks can deal with that by making predictions based on likelihood values instead of dealing with strict rules. Conati et al. (2002) demonstrated the ability of Bayesian networks to handle this inherent uncertainty. Another advantage of Bayesian networks is the fact that it represents causal relationships between the attributes in the network (Ball, 2001). This creates the possibility to explicitly model and test the relationships between measured signals from different modalities and their effect on the emotional state, as described by theories in psychophysiology (Conati, 2002). The following describes how we constructed a Bayesian Belief Network in order to equip our empathetic agent with the ability to recognize human emotions.

Methods A Bayesian Belief Network represents conditional dependencies among a set of attributes (Duda et al., 1973; Russell and Norvig, 1995). The structure of such a network can be graphically displayed as a directed acyclic graph in which each node represents an attribute and dependency relationships are indicated by the links in the network. A node is directly influenced by its preceding nodes (parent nodes) and it influences the set of nodes after it (child nodes). We need to specify two things to describe a Bayesian Belief Network B, the network structure B_s as represented by the directed acyclic graph and the parameters that represent the conditional probability distributions B_p, defining for every possible value of each attribute what its probability is given the values

of the parents of the node ($B = (B_s, B_p)$). Defined like this, the network gives a complete description of the domain. From the information in the network, a full joint probability distribution over an entry can be computed as shown in equation 4.4 where X_i is the i^{th} node in the network and x_i its value.

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|parents(X_i)) \tag{4.4}$$

With the use of an inference engine, the posterior probability distribution for a set of query variables given some observed variables (evidence) can be computed.

Both parts, the network structure and the conditional probability distributions can be learned from the data set. We used the Bayes Net Toolbox (BNT) (Murphy, 2001) in Matlab, which is a convenient open source toolbox, to train and test our Bayesian Network.

To learn the structure of our network from the data, we used a method based on Maximum Weight Spanning Tree searching (Chow and Liu, 1968). This method uses the notion of mutual information to find the dependence tree of maximum weight. Chow and Liu (1968) prove how this method functions as a maximum likelihood estimator of the tree.

After defining the network structure we need to learn the parameters of our network from the data to complete the network. A method that is used often for parameter learning is Maximum Likelihood estimation (Russell and Norvig, 1995). The problem with this method is that it needs the assumption that all proportions are equally likely a priori and also that it has difficulties with small data sets (Russell and Norvig, 1995). Since our data set is small and some emotions appear more often than others because of differences in movie length and elicitation success we decided to use another approach that is better suitable, Bayesian parameter updating as implemented in BNT. This approach places a hypothesis prior over the possible values of the parameters and this distribution is updated when new evidence is available.

The Junction Tree Inference Engine was used in our experiments to test the recognition accuracy in a leave-one-out cross validation procedure.

Four different static Bayesian Network classifiers were implemented and tested. Two conditions were altered to find the best way of modeling the data with Bayesian methods. One condition involved whether or not attribute selection was used to choose the best attributes and eliminate the rest. Attribute selection was performed using Information Gain Ranking in Weka (Witten and Frank, 2005). The other condition determines whether only one Bayesian Network was trained on the 5 class problem or 5 separate Bayesian Networks were trained on a two class problem for each emotion to determine the decision. The decisions of the combination of conditions that created the best recognition accuracy were combined with the FaceReader decisions to find out whether this would improve the performance. The two classifiers were combined using the sum rule (4.5) as described by Kittler et al. (1998). This rule combines the decisions of R different classifiers by summing the posterior probabilities $P(\omega_k|x_i)$ for each k of the total of m categories and choosing the category with the maximum combined posterior. This rule has been shown to outperform other classifier combination rules (Kittler et al., 1998).

$$\max_{k=1}^{m} \left[(1-R)P(\omega_k) + \sum_{i=1}^{R} P(\omega_k|x_i) \right] \tag{4.5}$$

Results The performance results of the four different Bayesian classifiers are reported as percentages of correctly classified instances in table 4.2.

	Attribute selection	All attributes
1 Bayes Net	41.2 %	28.1 %
5 Bayes Nets	43.5 %	30.3 %

Table 4.2: *Performance (in percentage correctly classified) of four Bayesian classifiers. Varying two conditions: in two cases attribute selection was used while in the other two all attributes were present in the data set; in two cases one Bayesian network was trained on the five-class problem while in the other two there were five Bayesian networks, one for each emotion as a two-class problem.*

Table 4.3 shows the confusion matrices of the classified emotions. The table shows for each emotion in percentages of the total amount of examples how the examples were classified. The bold font letters represent the emotions: **S** = Sadness, **D** = Disgust, **A** = Anger, **F** = Fear and **H** = Happiness. It gives a little bit more insight into the problem that is encountered when all attributes are used. These issues will be addressed further in the following paragraph about the discussion of these first results.

		Attribute selection					All attributes				
	Classified as →	**S**	**D**	**A**	**F**	**H**	**S**	**D**	**A**	**F**	**H**
1 BN	**S**	58	8	3	12	19	60	0	0	0	40
	D	40	16	4	12	28	42	0	0	0	58
	A	36	7	3	11	43	65	0	0	0	35
	F	41	10	2	21	26	67	0	0	0	33
	H	18	4	1	9	68	58	0	0	0	42
5 BN	**S**	63	0	1	11	25	73	0	0	0	27
	D	43	0	3	21	33	56	0	0	0	44
	A	29	0	2	17	52	73	0	0	0	27
	F	38	0	0	30	32	66	0	0	0	34
	H	18	0	0	9	73	63	0	0	0	37

Table 4.3: *Confusion matrices of Bayesian classifiers. This table shows for each emotion, in* ***percentages*** *of the total amount of examples, how the examples were classified. The bold font letters represent the emotions:* ***S*** *= Sadness,* ***D*** *= Disgust,* ***A*** *= Anger,* ***F*** *= Fear and* ***H*** *= Happiness. The boxes shaded in grey show the percentages of each emotion classified correctly.*

The above results were obtained using only the data from physiological signals. As mentioned before, we applied decision level fusion to combine the data from physiological signals with the facial expression recognizer. First, it might be interesting to see how well the facial expression recognizer performs on its

own. Unfortunately this performance is very poor, only 13.0 % is classified correctly by the FaceReader. Table 4.4 shows the confusion matrix.

Classified as →	**S**	**D**	**A**	**F**	**H**
S	4	11	82	3	0
D	4	34	49	0	13
A	8	9	83	0	0
F	20	12	66	0	2
H	16	11	65	0	8

Table 4.4: *Confusion matrix of FaceReader performance. This table shows for each emotion, in* ***percentages*** *of the total amount of examples in which FaceReader could detect a face, how the examples were classified. The bold font letters represent the emotions:* ***S*** *= Sadness,* ***D*** *= Disgust,* ***A*** *= Anger,* ***F*** *= Fear and* ***H*** *= Happiness. The boxes shaded in grey show the percentages of each emotion classified correctly.*

This is of course not very promising for the performance of the combined recognizer, but it has been shown that the combination of several modalities that have difficulties that vary in their nature can still have a complementing and improving power (Sharma et al., 1998). Therefore it seemed useful to try combining both classifiers. The performance of the combined classifiers is 23.5 % correct, which is higher than the performance of the FaceReader alone, but lower than the recognition result that can be obtained by using only the physiological signals. Table 4.5 shows the confusion matrix of this combined classification. A closer look at the FaceReader data reveals that in 39.6 % of the total video length, the face was not even found or the software was unable to create the face model. This explains why the recognition performance of the face reader alone is even below chance level. The confusion matrix shows the performance when only those cases are taken into account where the FaceReader was able to make a decision. In this case 21.5 % was correctly classified, only slightly above chance level. Furthermore, in the cases where it did find the face, 71.7 % was classified as Anger which indicates a strong bias towards this emotion. This also follows from the confusion matrix.

Classified as →	**S**	**D**	**A**	**F**	**H**
S	19	9	64	1	7
D	26	26	36	0	12
A	32	9	57	1	1
F	39	7	39	0	15
H	42	6	28	0	24

Table 4.5: *Confusion matrix of classifier that combines facial expression data with physiological signals. This table shows for each emotion, in* ***percentages*** *of the total amount of examples, how the examples were classified. The bold font letters represent the emotions:* ***S*** *= Sadness,* ***D*** *= Disgust,* ***A*** *= Anger,* ***F*** *= Fear and* ***H*** *= Happiness. The boxes shaded in grey show the percentages of each emotion classified correctly.*

Discussion In contrast to the high expectations the affective computing community has for Bayesian methods, the overall results are somewhat less than expected. These result do indicate that attribute selection can make a big difference in the performance and that the use of five separate Bayesian Networks instead of one increases the performance somewhat, but not spectacularly.

The main reason for the poor overall recognition performance of the Bayesian classifiers seems to be the big influence of the prior probabilities. The data set is asymmetric in the sense that there are a lot more examples in the set of certain emotions compared to others. This is due to the fact that some have been elicited successfully more often than others and because the movie clips did not have the same duration, so some sequences of feature vectors are longer than others. As follows from the confusion matrices, most instances will be classified as either Happiness or Sadness which are the two emotions that have the most examples in the data set and the highest prior probabilities. Sometimes, when the data is quite ambiguous and noisy, the posterior probability distribution ends up not differing too much from the prior probability distribution (Kittler et al., 1998). This could be the case in our experiments, but before drawing this conclusion more classification methods are applied to the data set in the next sections.

Considering this influence of the prior probabilities in case of ambiguous data, it is not so surprising that this effect is much stronger when all attributes are used. In the case of attribute selection, the personality data is completely excluded which influences the performance positively because actually this data is very ambiguous since it is constant for every person in the data set. It creates the same observation for every emotion condition for each test subject, which makes it very difficult to make a useful classification using this information. Therefore, from now on in the next section we will only consider classifiers using the attribute selection that was used above.

The disappointing result that we encountered with the FaceReader can best be explained by our experimental set up and the placement of the camera. The creators of FaceReader recommend to place the camera right in front of the screen which makes it view the face as much from the front as possible. In our experiment, the test subjects had to watch movies on the screen, so we decided to place the camera on top of the screen. This might explain why there is such a big bias towards Anger because the faces are viewed a little bit from above which often makes neutral faces appear as angry faces. Also, the expressiveness of the faces during the recordings was quite limited. The facial expressions were very subtle and most of the time the faces were neutral. Another reason for the fact that the software had difficulties finding and modeling the faces could be the lighting conditions. The FaceReader creators give very strict guidelines for the placement of light sources but unfortunately we did not have the possibility to adjust the light sources in the lab that was available to us for doing the experiments. Since the use of the facial expression data only lowers the performance of the classifiers, we decide not to use it in the experiments described below.

4.2.2 Dynamic Bayesian modelling

Another trend in the affective computing community involves the awareness that affect should be treated as a dynamic, temporal system (Pantic and Rothkrantz, 2003). Emotions evolve over time and the affective state that a person is in at

a certain time instance influences that of the next time instance. Trying to interpret a behavioral change in a time instance can also lead to mistakes. A frown of the eyebrows for instance can be interpreted both as anger and as confusion (Pantic and Rothkrantz, 2003). An expression of surprise followed by a smile also reflects a very different affective state than when it is followed by fear. Taking the temporal pattern of behavioral changes into account could prevent these ambiguities (Pantic and Rothkrantz, 2003). Using time sequence data is also expected to make the recognition results more robust since the nature of muscular action and physiological reactions make certain transitions from one state into the other less likely or even impossible. This knowledge can aid the recognition process and causes the system to not be completely helpless in the case of inaccurate measurements or missing data. The following describes how we attempted to extend our Bayesian Belief Network to be dynamic.

Methods Dynamic Bayesian Networks differ from static Bayesian Belief Networks in their ability to process temporal data. Instead of feeding just one feature vector into the network as evidence, a sequence of feature vectors can be used so that the decision making in a certain time slice can also benefit from observations done in previous time slices. To define a Dynamic Bayesian Network we do not only need to specify the intra slice structure (observation model), $P(E_t|X_t)$ (with E_t being the evidence) but we also need to create an inter slice topology (transition model), $P(X_{t+1}|X_t)$. The network parameters are defined by the conditional probability distributions over two time slices, called a two slice temporal Bayes Net (2TBN). Again, the Bayes Net Toolbox (Murphy, 2001) was used to implement and test our Dynamic Bayesian Network.

Results The performance of the Dynamic Bayesian classifiers was tested only with attribute selection this time but again while using either one or five separate networks. In the case of one Dynamic Bayesian network, 31.9 % was correctly classified and in the case of 5 networks the performance was 29.0 %. Table 4.6 shows the confusion matrices for both experiments. Interesting is the fact that in this case one network seems to perform better than the combination of five separate ones as opposed to the case with the Static Bayesian Networks. Overall the performance of the Static classifiers were better than the Dynamic ones.

Discussion As opposed to the expectations deduced from the recent developments in affective computing, our results show that on our data the Static methods work better than the Dynamic ones. Due to the way the data set is constructed and the elicitation method that has been used, it is actually not very surprising that the performance would not increase spectacularly because the sequences of feature vectors are very short for every emotion and the way the different emotions follow one another is not very natural. It is not expected to be very common for people to switch from Fear to Happiness to Anger in such a short time. Still this can not explain why the results are much worse than in the Static case. One explanation could be that, because the decisions are influenced by earlier decisions, if the classification is wrong in the beginning of the sequence it increases the probability also to be wrong during the rest of the sequence. Since the sequences are so short, this might mean that the system does not even get the chance to correct itself before the emotion is over.

	Classified as →	**S**	**D**	**A**	**F**	**H**
1 BN	**S**	34	8	15	32	11
	D	21	15	0	44	20
	A	28	0	12	31	29
	F	35	10	0	16	39
	H	10	9	6	17	58
5 BN	**S**	29	16	16	28	11
	D	36	22	0	22	20
	A	37	3	21	8	31
	F	33	26	7	6	28
	H	19	5	8	18	50

Table 4.6: *Confusion matrices of Dynamic Bayesian classifiers, both in the case of one network that models the five-class problem and in the case of five networks combined, one for each emotion modelling a two-class problem. This table shows for each emotion, in* ***percentages*** *of the total amount of examples, how the examples were classified. The bold font letters represent the emotions:* ***S*** *= Sadness,* ***D*** *= Disgust,* ***A*** *= Anger,* ***F*** *= Fear and* ***H*** *= Happiness. The boxes shaded in grey show the percentages of each emotion classified correctly.*

Another issue could be the size of the data set. 25 test subjects is not much and in the Dynamic case an example is the complete sequence for one emotion for each subject. This means that there could be a maximum of 125 examples, but given the fact that we had to exclude a lot of data, we only had 81 examples left. In the Static case, every 40 seconds of the recording a new example was created, so the Dynamic experiment might have suffered more from the small size of the data set.

4.2.3 k-Nearest Neighbours

In order to compare the results of the Bayesian methods to the performance of a non-Bayesian classifiers on our data, we used the k-Nearest Neighbour approach as another affect recognizer. The k-Nearest Neighbour algorithm compares every new test instance to the training instances with a distance measure to define its k nearest neighbours. The test instance will be classified in the category that is most common among its nearest neighbours.

Orange (Demsar et al., 2004), a machine learning toolbox, was used to run the k-Nearest Neighbour classifier on our data. We used k = 6 neighbours and the Euclidian Distance measure to compute these nearest neighbours. With this Nearest Neighbour classifier 44.4 % of the examples was classified correctly. Table 4.7 shows the confusion matrix of these results.

The performance of the k-Nearest Neighbour classifier is comparable to that of the best Static Bayesian classifier on our data. The confusion matrix shows that this method also suffers from the unbalanced division of categories in the data set, but the effect is less severe than with the Bayesian Methods. Overall, the fact that the performance is about the same with this non-Bayesian method strengthens the surmise that it is probably due to the nature of the data set that a higher performance can not be reached.

Classified as →	**S**	**D**	**A**	**F**	**H**
S	51	10	12	12	15
D	20	30	11	15	25
A	20	6	35	13	26
F	22	6	4	44	23
H	15	9	10	16	49

Table 4.7: *Confusion matrix of k-Nearest Neighbour classification results. This table shows for each emotion, in* ***percentages*** *of the total amount of examples, how the examples were classified. The bold font letters represent the emotions:* ***S*** *= Sadness,* ***D*** *= Disgust,* ***A*** *= Anger,* ***F*** *= Fear and* ***H*** *= Happiness. The boxes shaded in grey show the percentages of each emotion classified correctly.*

4.2.4 Alternative approach with continuous emotion representation

As discussed in the background section, not all psychologists agree that emotions can be categorized in a discrete manner. Even though many affective computing researchers are aware of the problems with the discrete approach, it still used a lot more often than dimensional approaches. This might be due, as it was in our case, to the fact that discrete emotions are easier to work with and much more intuitive for participants that are asked to self report their emotional state. We are much more used to talk about our emotions in terms of emotion-words than in terms of dimensions like valence and arousal. In addition to this, when information from several modalities has to be combined and there are constraints on the way emotions can be elicited, it is sometimes more convenient to use discrete emotions. In our case for instance, we used the FaceReader to recognize facial expressions and this tool used discrete categories. The movies we used for emotion elicitation were also selected to elicit specific discrete emotion categories. Realizing that these parts need to be combined, the choice to use a discrete categorization is made easily. Peter and Herbon (2006) propose to abandon the habit of naming emotions with words completely and they provide guidelines about how to structure emotions as a dimensional representation for the use in human-machine interaction. Labelling emotions can be problematic because the category borders are blurry and the word Anger for instance can describe many different emotional states. The method as described by Peter and Herbon (2006) avoids these problems because it abandons labelling emotions with words. It consists of the following four steps:

- Step 1: Eliciting emotions while measuring physiological signals and asking test subjects to self report in a way that can be translated into a dimensional structure.
- Step 2: Assigning the physiological measurements to the related ratings.
- Step 3: Group emotions into clusters with similar physiology and place in dimensional structure.
- Step 4: Identify characteristic patterns in physiology for each cluster.

Using the emotion ratings that our test subjects did with the Emotion Wheel during our experiments, we attempted to use this strategy to find a mapping from physiological signals to Scherer's appraisal dimensions. As mentioned before, the Emotion Wheel is a tool that allows users to do self report in a way that is intuitive for them with the use of emotion words, but the underlying structure translates into two important appraisal dimensions as described in the component process theory (Scherer, 2005).

The first step of the procedure had already been included in our experimental set up. We elicited emotions through movie fragments, measured physiological signals and asked participants to self report using the Emotion Wheel. The second step was an easy one as well in which we assigned the physiological measurements during a movie fragment to the Emotion Wheel rating for that movie. For the second step we implemented the K-means clustering algorithm (Duda et al., 1973) which tries to find the centers of clusters that are naturally present in the data. Then with this new class labelling that is not associated with emotion words, we trained and tested Static Bayesian Networks again with attribute selection. In the case of one Bayesian Network the performance this time was 58.7 % classified correctly and in the case of six Bayesian Networks it was 60.0 %.

4.3 Summary

In this chapter, the design and implementation of an affect recognizer has been described. First, a data set had to be created which has been used to train and test several classifiers. We used a Static Bayesian Network, a Dynamic Bayesian Network, a simple k-Nearest Neighbour classifier and an alternative approach in which a continuous instead of a discrete emotion representation was used. The alternative approach was able to reach the best performance of all the classifiers on our data. The performance results of the Static Bayesian Network and the k-Nearest Neighbours classifier were almost the same and the Dynamic Bayesian Network performed the worst. The results are summarized in table 4.8. Some possible explanations for these results have already been mentioned in the individual discussion sections. These issues will be discussed further in chapter 6.

	SBN	**DBN**	**kNN**	**PH**
Performance	43.5 %	31.9 %	44.4 %	60.0 %

Table 4.8: *Summary of the performance (in percentage correctly classified) of all classifiers that have been applied to the data. The percentage represents the maximum performance that was reached for each method.* ***SBN*** *= Static Bayesian Network,* ***DBN*** *= Dynamic Bayesian Network,* ***kNN*** *= k-Nearest Neighbour,* ***PH*** *= alternative approach based on (Peter and Herbon, 2006).*

Chapter 5

Affect generation

Once a machine is able to recognize the affective state of a human user, it is useful if the machine can use this information to adapt its behaviour appropriately to the user. As mentioned before, we do not focus on a specific application area but we want the feedback to consist of a basic empathetic response which shows the users that the machine is aware of their affective states. A first step towards realizing this empathetic response is to mirror the affective state of the user. This chapter describes the design and implementation of affect generation in an anthropomorphic avatar.

As mentioned before, the human face and the use of facial expressions are of a very big importance in human-human interaction. It has already been shown that in several application areas such as education (Lester et al., 1997), answering questionnaires (Walker et al., 1994), e-therapy (Grolleman et al., 2006) or robotics (Bruce et al., 2002), the human-machine interaction can benefit from the presence of a virtual face. In order to improve human-machine interaction and to create more pleasant usability, an anthropomorphic graphical avatar could provide the appropriate empathetic feedback by showing affective expressions in a natural and believable way.

The question of how to create anthropomorphic avatars that are both engaging and believable and the impact of these avatars on human emotions and user experience is still being studied (Prendinger et al., 2004). There are many pitfalls which should be carefully avoided, such as the issue of the *uncanny valley* as described before. The best examples of believable animated virtual expressions may be found in the gaming and movie industry. Modelling teams spend a large amount of time to create animations that are certainly engaging and become more and more believable. Some disadvantages of the approaches applied in this area is that it is very time-consuming and expensive. At Pixar[1] for instance, a complete team of animators sometimes work weeks to complete 3 seconds of a movie to make it as believable as possible. Moreover, often actors are used to animate expressions and movements from motion capture. In research, unfortunately, there is not as much time, money and manpower available as there is in the entertainment industry, so it is unreasonable to assume we can reach the level of believability and engagement that we see in the animated movies and games, but luckily some work has been done and several standards

[1]Pixar Animation Studios is one of the most successful computer animated movie production companies, California, USA

and scripting languages have been developed for animating facial expressions on anthropomorphic avatars such as the MPEG-4 FAP (Facial Animation Parameters) standard (Ostermann, 1998) and the CML (Character Markup Language) (Arafa and Mamdani, 2003). This section describes an attempt to model 11 psychologically grounded facial expressions for an avatar.

5.1 Psychologically grounded facial expressions

Several building blocks are needed in order to create psychologically grounded facial expressions for an avatar. First we need a tool that can be used to create and animate expressive avatars. Second, a theory about emotions is needed that defines exactly how emotions are linked to facial expressions and third we need a way to translate the theory into paramaters that make sense to the avatar tool. The tool that was used here is called Haptek[2]. Haptek's People Putty can be used to create an avatar and a scripting language allows for easy animation of the face. A theory that seems very suitable to guide the implementation of the expressions is the component process theory as described by Scherer (2001) which has been mentioned before in the background section. This theory explains emotions and the ways in which they can be differentiated from one another as the result of a sequence of evaluation checks (appraisal). These checks help people to asses for instance how new, desirable or controllable a certain situation is for them and this is what defines their affective state. The set of four main appraisal domains of sequential evaluation checks (Scherer, 2001) is as follows (from lower to higher level and occurring in this order):

- Relevance - *Does this event directly affect me or my loved ones?*
- Implications - *What are the consequences?*
- Coping Potential - *How well can I adjust to these consequences?*
- Norm Compatibility - *How important is the event in terms of my norms and values?*

Scherer and Ellgring (2007) describe with quite a lot of detail how exactly these sequential evaluation checks are linked with our physiological features such as facial expressions, vocal intonation and the autonomic nervous system. To predict facial expressions in the context of certain sequential evaluation checks, Scherer and Ellgring (2007) created guidelines that describe for different emotions which Action Units (Ekman and Friesen, 1978) are active during each stage in the sequence of evaluation checks. This is what makes Scherer's theory so suitable for this task. Figure 5.1 shows an example of how this theory would predict a sequence of facial expressions.

The work that is presented in this section is an extension on what has previously been done by Paleari and Lisetti (2006a). Paleari and Lisetti (2006a) created a translation of Action Units into Haptek parameters which made it possible to apply Scherer's guidelines in the animation of facial expressions on Haptek avatars. Paleari and Lisetti (2006a) implemented only five different emotions while Scherer and Ellgring (2007) describe facial expression predictions of more emotion categories. Following the same strategy, we implemented

[2]http://www.haptek.com/

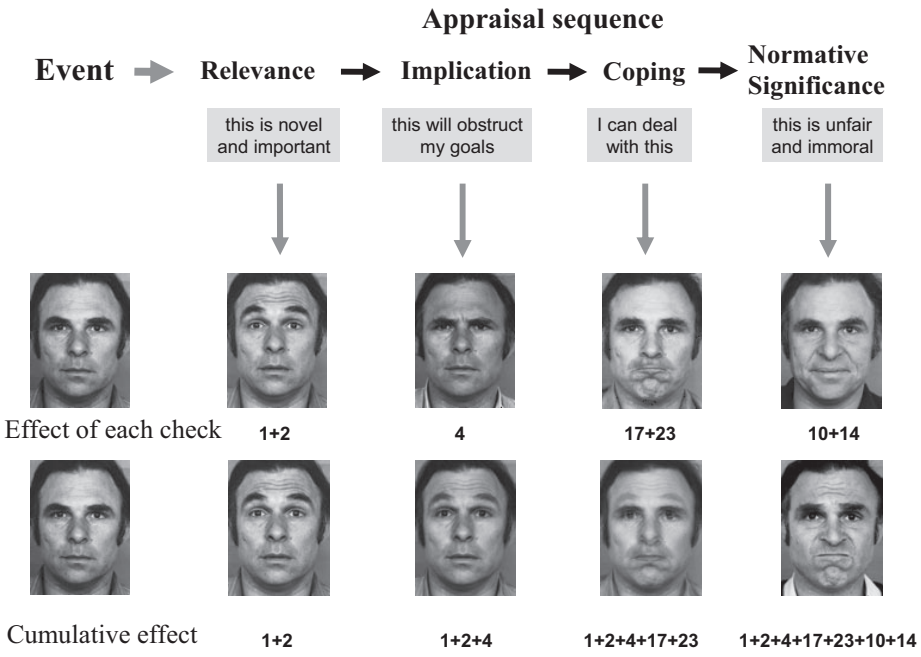


Figure 5.1: *From Scherer and Ellgring (2007), example prediction of facial expression sequence*

seven more emotions on an avatar and conducted a small user study to test all twelve of them together for recognizability and naturalness.

Methods Haptek avatars can be controlled by a hypertext scripting language. In these scripts, Haptek switches can be activated which are used to control facial parameters on a lower level. Using the translation from Action Units to Haptek parameters, Paleari and Lisetti (2006a) created switches that could be used to animate the following emotions: Happiness, Anger, Sadness, Fear and Disgust. We expanded this set by adding Shame, Pride, Interest, Surprise, Boredom, Contempt and Despair. Scherer and Ellgring (2007) describe for every emotion how it responds to each evaluation check and for each check they define how it influences the facial Action Unit. When we combine these conversions, we can create a table that defines for every emotion, for each check which Action Units are active. Table 5.1 shows for all 12 emotions our interpretation of the appropriate Action Unit sequences. This information could be used as input for creating the Haptek switches in which the Action Units could be represented by Haptek facial parameters in accordance with the translation created by Paleari and Lisetti (2006a).

Appraisal domain	Check	Happiness	Anger	Sadness	Disgust
Relevance	Novelty	1, 2, 5	1, 2, 25	4, 7, 17, 23	4, 7, 26, 38
	Intrinsic pleasantness	12, 25		1, 15, 41, 64	10, 15, 17, 26, 38
Implications	Expectation		4, 7		4, 7
	Goal attainment	12, 25	17, 23, 25		9, 10, 15, 17, 26, 38
Coping potential	Power/control		17, 24, 4, 7	26, 4, 1, 15, 41, 64	23, 9, 15, 17, 26, 38
Norm compatibility	External standards		10		10
	Internal standards				

Appraisal domain	Check	Interest	Boredom	Shame	Pride
Relevance	Novelty	1, 2	1, 2	1, 2, 4	1, 2, 13
	Intrinsic pleasantness	12, 25, 5	25, 41, 61	25	12, 25, 15
Implications	Expectation				
	Goal attainment	12, 25, 53, 54	25, 14, 62	5, 24, 62, 63, 64	6, 12, 25, 17, 23
Coping potential	Power/control	26		25, 23, 51, 52	6, 17, 24
Norm compatibility	External standards				
	Internal standards		14	14	

Appraisal domain	Check	Fear	Despair	Contempt	Surprise
Relevance	Novelty	1, 2, 5	1, 4, 7	4, 7	1, 2, 4, 7
	Intrinsic pleasantness	25, 1, 2, 5	11, 25, 43, 15, 64	10, 15, 25, 9	10, 15, 5
Implications	Expectation		4, 7, 28, 15, 1, 24	4, 7	4, 7
	Goal attainment	7, 23, 17, 25, 1, 2	6, 25, 17, 51, 52, 64, 15	25, 17, 10, 64	
Coping potential	Power/control	1, 2, 5, 26, 23, 17	26, 1, 7, 15	26	20, 26
Norm compatibility	External standards			10	
	Internal standards				

Table 5.1: *Action Unit activation in the order of evaluation checks as deduced from the component process theory of emotions. This table is adapted from the one in (Scherer and Ellgring, 2007) where these sequences are displayed for a few emotions. This table shows our interpretation of the translation from sequential evaluation checks to Action Unit activation in facial expressions for twelve different emotions. The facial expressions on our avatar were modelled with the help of this interpretation.*

Figure 5.2 shows a snapshot for each of the 12 emotions. In some cases it was needed to fine tune the expressions by adding or removing parameters and by adjusting parameter intensities. A problem that had already been encountered by Paleari and Lisetti (2006a) was that the theory is too unclear about intensity levels and timing issues, so how long exactly one evaluation check lasts and how it transmits to the next had to be determined ad hoc. In addition to this, the theory does not take Action Units for head movements (looking up/down/left/right) into account, but these movements are quite important. The chin is lifted when we are proud for instance and it points down when we are sad. Therefore we also added Action Units for these movements.

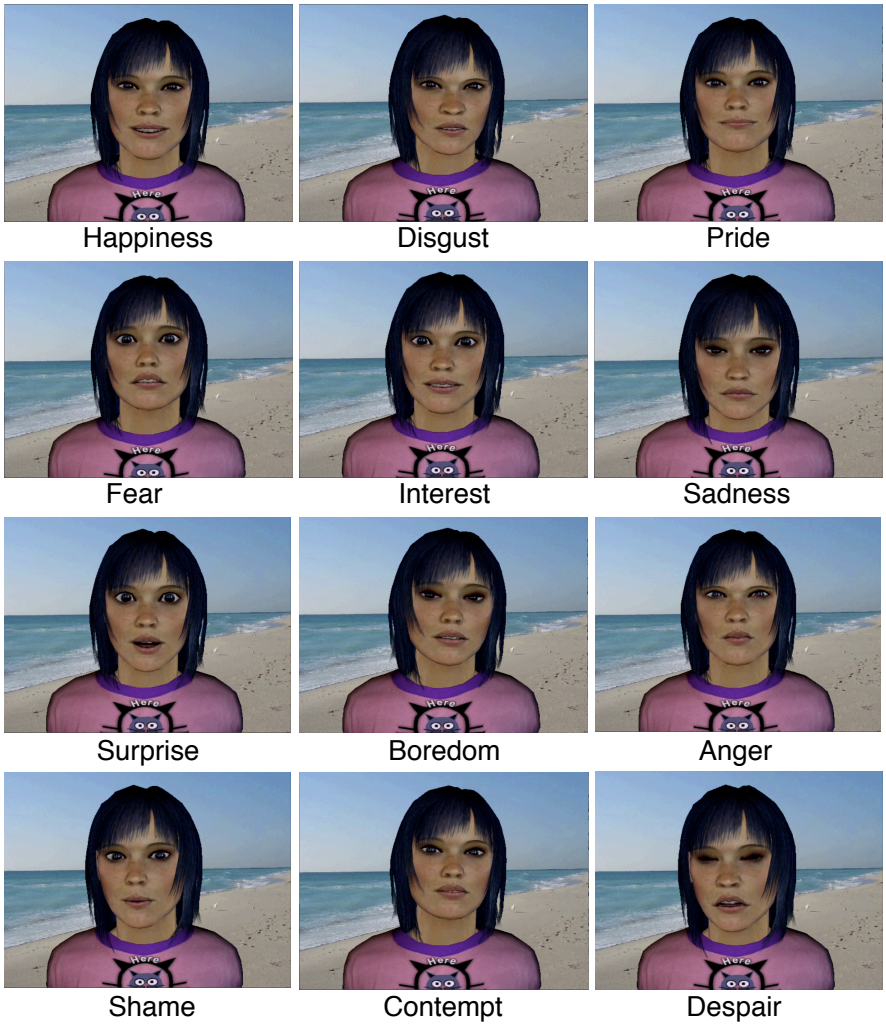


Figure 5.2: *Snapshots of the 12 implemented emotions on the Haptek avatar*

A small user study was conducted to test the new avatar expressions for recognizability and naturalness. Test subjects were asked to watch a movie for each emotion animation. Then they were asked to indicate for each movie

which emotion they recognized in a multiple choice question with the choices: Anger, Happiness, Surprise, Pride, Fear, Disgust, Interest, Contempt, Shame, Boredom, Sadness, Despair, Neutral, No idea. In addition to this they had to rate the naturalness of the expression on a scale from 1 to 5: 1 = Not at all natural, 5 = Very natural. 22 persons (11 male and 11 female) participated in this experiment varying in age from 21 to 62.

Results Table 5.2 shows the recognition scores in percentage for each animated emotion defining how many people recognized this emotion correctly as well as the average ratings for naturalness. The total performance of all participants over the complete set of emotions is 44 %. The average believability rating over all emotions is 3.05. In 8.3 % of the total amount of ratings people had no idea which emotion was expressed.

Emotion	Recognition score	Believability	SD
Happiness	50.0 %	3.20	1.04
Disgust	36.4 %	3.25	1.04
Pride	36.4 %	3.38	1.19
Fear	40.9 %	2.89	0.78
Interest	54.5 %	2.92	1.08
Sadness	54.5 %	3.75	0.97
Surprise	81.8 %	3.44	0.92
Boredom	50.0 %	3.00	1.26
Anger	63.6 %	3.23	1.17
Shame	13.6 %	3.00	1.00
Contempt	18.2 %	2.50	0.58
Despair	22.7 %	2.00	0.71

Table 5.2: *Recognition scores for avatar expressions, average believability (naturalness) ratings and Standard Deviation of the believability ratings.*

In order to give a little bit more insight into the difficulties that were encountered with this task, table 5.3 shows a confusion matrix of how the emotions were recognized by the participants.

Discussion As reported by Paleari and Lisetti (2006a), when participants were asked to recognize the five emotions that they had animated, the performance was 94 %. This means that expanding the set of emotions to the amount of twelve resulted in a huge performance drop. Only 44 % was recognized correctly in this case. Of course the task is a lot more difficult when there are more than twice as many categories to choose from. There are not only more categories but the expansion also introduced more emotions that are very similar and are easily confused. Disgust and Contempt for instance are very hard to distinguish as also follows from the confusion matrix. The differences between them are only subtle. The same problem counts for Despair with Sadness, which are very similar emotions as well and apparently Interest is confused often with Shame and Pride. Therefore it is not surprising that the performance drops. People would have difficulties distinguishing these emotions on human faces as well. In fact humans are not perfect emotion recognizers, they make mistakes

↗	Ha	Di	Pr	Fe	In	Sa	Su	Bo	An	Sh	Co	De	Ne	No
Ha	50	0	0	0	18	0	32	0	0	0	0	0	0	0
Di	0	36	0	5	14	0	0	0	23	0	4	4	0	14
Pr	5	0	36	0	23	0	5	9	0	0	9	0	4	9
Fe	0	0	0	41	9	0	14	4	0	0	5	14	4	9
In	0	0	0	0	55	0	18	5	0	0	0	0	18	4
Sa	0	0	0	0	0	55	0	18	5	14	4	0	0	4
Su	5	5	0	0	0	4	82	0	0	0	4	0	0	0
Bo	0	4	4	0	0	5	5	50	0	9	14	0	5	4
An	0	18	0	0	0	0	0	0	64	0	9	0	0	9
Sh	4	0	0	0	36	5	23	0	0	14	0	0	0	18
Co	0	18	0	9	14	0	0	0	0	4	18	0	14	23
De	0	9	0	0	0	50	0	0	5	0	5	23	4	4

Table 5.3: *Confusion matrix of avatar expression recognition by participants. This table shows in* ***percentages*** *how often an expression was recognized as one of the 14 choices in the experiment. The cells that are shaded in grey indicate the percentages that each expression was classified correctly.* ***Ha*** *= happiness,* ***Di*** *= disgust,* ***Pr*** *= pride,* ***Fe*** *= fear,* ***In*** *= interest,* ***Sa*** *= sadness,* ***Su*** *= surprise,* ***Bo*** *= boredom,* ***An*** *= anger,* ***Sh*** *= shame,* ***Co*** *= contempt,* ***De*** *= despair,* ***Ne*** *= neutral,* ***No*** *= no idea.*

as well and it is very clear for instance that females perform a lot better at this task than men (Biele and Grabowska, 2006).

As mentioned before, the component process theory is unclear about the exact timing of the evaluation checks and how the transitions from one check to another is supposed to be realized. It might be beneficial for the recognizability and believability of the animated expressions if these issues could be modelled based on psychologically grounded ideas as well.

One of the participants, after finishing the experiment, commented that the task was so hard because the right context was missing in the movies. In a setting like this on the beach with nice weather you would not expect someone to show an expression of despair. It is also very well imaginable that anger, disgust and contempt are much more easily distinguishable when the context is known. It is therefore expected that people have less difficulties recognizing animated expression when they are shown in a real human-machine interaction context, because of their expectations and experience.

A useful extension to this set of expressions might be the use of full torso avatars in which hand gestures can be included as well. Some emotions may rely more on hand gestures than others, but in some cases certain gestures are very typical for the emotions. In the case of shame for instance, it is very common that people try to hide behind their hands and in the case of boredom people often rest their head on their hands or arms. Despair often makes people grab their head. Using only the face to animate the expressions, it is impossible to incorporate these cues.

Chapter 6

Overall discussion and conclusion

In the previous chapters, the design and implementation of two important building blocks for creating affective human-machine interaction systems have been discussed. The first building block involved the training and testing of several emotion recognizers. Based on recommendations and expectations from the affective computing community, Static and Dynamic Bayesian Networks were used to create affect recognizers that have been trained and tested on a data set we created with human test subjects. In addition to this, a k-Nearest Neighbour approach was used as well as an alternative approach, in which we used the method that was described by Peter and Herbon (2006) to classify the data into categories that were based on a continuous emotion representation. The highest performance that was reached with Static Bayesian Networks was 43.5 % classified correctly. Dynamic Bayesian Networks were able to classify 31.9 % correctly, the performance of k-Nearest Neighbours was 44.4 % and the alternative approach classified 60.0 % correctly. In order to be used in a serious application where affect needs to be recognized in a reliable way, these results might not suffice. In the next section a diagnosis will be given in which we try to find possible explanations for these results which will facilitate the design of future experiments and implementations. Before this, we will review our results in relation to the appropriate expectations that one should have from systems like the one we attempted to create, which might yield a more positive evaluation of the performance achieved in the present research. One of the pitfalls in many machine learning application areas is that people expect the machines to work perfectly. Speech recognizers, hand writing recognizers and also emotion recognizers are not allowed to make any mistakes. From the perspective of a user this is very understandable. People do not want to buy or use systems that only partly work. From the research perspective, however, it is unrealistic to expect a 100 % performance on these tasks that even humans can not do perfectly. As an example, Ekman and O'Sullivan (1991) once investigated human abilities to detect whether persons on video recordings were lying or telling the truth about their emotional state. As many as 509 participants were tested, among which even psychiatrists, judges and people from U.S. Secret Service were present and they were hardly able to perform better than chance. Also, considering existing

systems and previous work that has been done, performances of about 80 % / 85 % can be achieved on a two-class problem, for instance when distinguishing stress from non-stress (Barreto et al., 2007) or interest from non-interest (Kapoor and Picard, 2005). Trying to classify five emotions is of course much more challenging and in this view a performance of twice or three times the expected performance by chance (20 % in this case) can actually be considered a good start towards a system that resembles human performance.

The second building block involved the design and implementation of psychologically grounded facial expressions on an anthropomorphic avatar. Using the component process theory of emotions (Scherer and Ellgring, 2007), twelve facial expressions were created for an avatar and these were tested for recognizability and believability in a small user study. Participants in this study were able to recognize the right emotion in 44 % of the cases. Surprise and Anger were recognized the best with results of 82 % and 64 % respectively. Again, from the perspective of applications, the obtained results might not be sufficient.

In the next section we will try to diagnose where the problems may lie in the whole process from data collection to recognition and generation results in order to provide a list of recommendations for future research that could improve the results that we found. This chapter will end with a final conclusion.

6.1 Diagnosis and recommendations

6.1.1 Affect recognition

Like we discussed before, we suspect that the attributes in our data set have become too ambiguous to be able to distinguish between five different emotions. There may be some possible causes for these results, therefore we will discuss in this section why and how it could happen that the feature vectors are this ambiguous.

One possible reason for the ambiguities in the data might be a failure with the emotion elicitation process. Even though we only use the data in which the self-report of the participants matched the intended emotion for the movie fragment, it could still have been the case that the emotions were so subtle that they were invisible in the physiology of the person. Also, the self report could simply be wrong. Maybe people self reported unconsciously what they thought they were supposed to feel and not what they actually felt. This would be in accordance with the fact that the alternative approach produced better results because this indicates that a better labelling for the values of the signals in the data set is possible. On the other hand, the movies have been selected, by previous researchers as well, because they are known to be suitable for eliciting these emotions.

As a second concern, we could criticize the feature extraction proces. The set of features that we computed from the signals have been proven to work for the distinction of stress from non-stress (Barreto et al., 2007), but our case is different. It may be that there are other characteristics in the signals that we overlooked and that would have provided more capabilities of distinguishing between different emotions. Also, we computed the features over much smaller intervals of the signals which could have altered the quality of the feature vectors. An example of a problem that could have arisen from this concerns the

computation of the Low Frequency/High Frequency ratio of the Blood Volume Pressure signal. Since we assessed segments that are much shorter of length than those that were assessed by Barreto et al. (2007), the occurrences of Low Frequency band events will be much less in one segment and could sometimes even be too slow to be detected at all.

A third problem might have been the discretization step on the data set. The method that we used has been tested and proven to work well on a range of different data sets (Holte, 1993). Therefore, it is not in the first place the algorithm for doing the discretization that should be doubted, but it might be the discretization proces itself that lowers the performance. When making a data set discrete, a lot of detailed information is simplified extensively which basically means that information is lost.

Another problem may lie with the recognizers. The Bayesian methods did perform below our expectations and are maybe not as suitable for this task as the high expectations from the affective computing community would predict, but the performance is not much better with several other classifiers on this specific data set. These algorithms are common methods used in machine learning, so it is very unlikely that the recognizers are a problem. Especially because there are several reasons to believe that our data set could have been constructed better.

Another reason for these results might be more problematic for the affective computing community. The physiological signals may not contain enough discriminating information to distinguish between five or more emotions. The fact that no classifier was able to recognize meaningful patterns in the physiological signals of our data set creates doubts about the correlation there is between these signals and the group of emotions. Certain correlations have certainly been found in the past, such as the correlation between stress and these physiological signals as has been found by Barreto et al. (2007). Of course it is a much more challenging task to classify a group of specific emotions and there was no guarantee that the necessary information would really be present in the measurements. It is sometimes doubted whether it is possible to find an exact mapping from signals to specific emotions. Findings such as the lack of correlation in signals that can be found for positive emotions (Peter and Herbon, 2006) are still problematic in the task of finding a structure and representation that can be used in computing. But the fact that we did find better results when we used the alternative approach gives hope that there is at least some structure in the data that can be found. Moreover, we still think that the use of physiological signals in combination with facial expression data can be much more successful. Unfortunately we have not been able to prove this because the FaceReader did not work properly on our video recordings, but with just two sensors the physiological signals might have suffered from a lack of discriminability on this five-class problem that could have been solved with the addition of the facial expression recognizer or other modalities.

A last, and probably one of the most severe problems is the small size and the unbalanced composition of the data set. Only 25 test subjects participated in our data collection experiment and a lot of data from these participants had to be excluded due to unsuccessful elicitation or problems with the recordings. Working with small sample sizes increases the difficulty of finding the right generalization for the underlying distribution. The difference in elicitation success and duration of movie clips for the emotions also caused the data set to be very unbalanced. It contained for instance more than twice as many examples of

Happy than of Disgust.

6.1.2 Recommendations for affect recognition

It is unclear what it is exactly that caused our data to be too ambiguous to reliably classify five emotions in our experiments, but certainly there is a lot that can be improved in future follow up experiments. Some recommendations for future work will be presented here.

Possible points for improvement in affect recognition:

More modalities - It has already been pointed out that the two sensors that we used for recording the physiological signals might not have contained enough discriminating ability to classify into five emotion categories. The fact that we could not use the FaceReader data caused that we could not test it in combination with other modalities but we still think that it is promising to combine the physiological signals with facial expression data and maybe even more modalities such as vocal intonation.

More participants - In the present research only 25 test subjects participated in the data collection. A lot of this data had to be excluded which resulted in a data set that was quite small. In a follow-up experiment, more participants should be invited so that the problem of not having enough data can be eliminated.

Other elicitation method - The way that we elicited the emotions in the experiment was one of the causes for the unfortunate composition of our data set. The stimuli should be of a longer duration so that dynamic methods can seriously be applied and they should have the same duration for all emotions in order to be able to create a balanced data set which contains the same amount of examples for each category.

Better use of personality data - It has been mentioned before that the way we used our personality data it only made the data set more ambiguous. If there is enough data so that there are several examples in the data set for each combination of personality traits, this data could be used to train several different personality-specific recognizers. In this way it would be possible to actually take the difference in physiological response between different personalities into account. In the case of Bayesian Networks for instance, the conditional probability distributions will then be adjusted to fit to the personality traits. This could for instance result in a higher prior probability for negative emotions if the person is neurotic and for Anger if the person is aggressive. In the present research there was not enough data to do this.

6.1.3 Affect generation

The recognizability of the psychologically grounded facial expressions that were implemented on an anthropomorphic avatar was lower than we hoped. As already mentioned in chapter 4, one reason for this result could be the fact that it actually was a very difficult task. Not only because the examples contained many emotions that are very similar such as contempt, disgust and anger or

sadness and despair, but also because information about context and other cues such as hand gestures were missing. Moreover, the participants had to choose the right emotion out of twelve possibilities, which makes the expected performance if random guessing would be used 8.3 %. The amount of examples that were recognized correctly were clearly higher than this for every emotion involved.

Another problem with the implementation of the expressions may be the incompatibility with the psychological theory that was used. The component process theory (Scherer and Ellgring, 2007) seemed to be the best suitable in this context because of the detailed information it provides about the relationship of sequential evaluation checks to dynamic facial expressions. Still, the theory is not perfect because some information, for instance about the exact timing of the checks and the amount of overlap between checks, is missing.

6.1.4 Recommendations for affect generation

In order to be able to reach higher recognition scores in possible future follow-up experiments, there are some points that can be improved in our methodology. Some ideas for improvement will be listed here for the design and implementation of more recognizable and believable animated facial expressions on the avatar.

Possible points for improvement in affect generation:

More research on timing - The component process theory of emotions (Scherer and Ellgring, 2007) is very informative about the relation between emotions and facial muscle movements but an exact description of the timing and overlap between these movements is missing. More research on these issues could help to improve the believability and naturalness of the animated facial expressions.

Adding context - One of the participants in the small user study commented that the context might have had an influence on the difficulty of recognizing the emotions. The movies showed a girl on the beach which might have diminished the expectation that someone shows an expression of despair. When the context is present, our expectations could help us to recognize the expressions better. In the case of a real application, this context is of course automatically present, so it might be beneficial to do a user study in a real application setting.

Full torso or full body avatars - People do not only use their face to express emotions, but in combination with our facial expressions, our hands and posture also reveal a lot of information about how we feel. As mentioned before, it is very common that people try to hide behind their hands in the case of shame for instance, and in the case of boredom people often rest their head on their hands or arms. Despair often makes people grab their head. The distinction between several emotions that are very similar might become more clear when full torso or full body avatars are used.

6.2 Future

We might face a future in which our computers can sense, interpret and react to our affective states. A machine that knows when you are frustrated with it, can sense when you read an e-mail that touches you, can choose to play the music you are in the mood for and does many more things to let you know it knows how you feel. Is that a machine people need and are comfortable to work with? In the introduction we already pointed out that people constantly show their emotions to their computer, even though they are perfectly aware that it does not understand it at all if they yell at it and blame it for being slow or stupid. Somehow it gives us a feeling of relief to release these frustrations towards the computer, even though it does not listen to the complaint. Would it not be even better if it did?

In elderly care, robots will be companions that are involved in the everyday life of a person. They will fulfill a large social role and people will grow attached to their electronic pets. Research[1] has been done in an elderly care home where Aibo's were placed in the resident's homes as pets. The elderly persons that were given such an electronic pet not only became more healthy (they got more exercise, acted more social towards other residents and smiled more often), but they also developed feelings for their pet. When they were asked to return the Aibo at the end of the experiment, some of the elderly people became very sad. To the question whether they would buy their own Aibo, one of them answered: "No, it wouldn't be the same dog and this one has such a great personality".

Kismet is another robot that people apparently can not treat as a machine. In the experiment (Breazeal and Aryananda, 2002) where people were asked to talk to Kismet with varying intonation in their voice, people reported to feel very guilty for speaking with an angry voice to the robot and they felt bad for making Kismet sad. It has been said that this kind of behaviour reflects the satisfaction of an instinctive human need to ascribed emotions to people, animals and even machines (Picard, 1997). Even though they know very well that the machine consists of sensors, plastic and electronics, they interact with it like they would with living creatures. In this sense, affect-sensitive machines and robotic companions are expected to be able to greatly improve the social well-being of humans.

6.3 Final conclusion

This thesis provided an overview of the design considerations and many challenges that are encountered in the interdisciplinary field of affective computing. After having discussed all the possible improvements that can be done to the methods that have been described in this thesis, it should be clear that a lot of work still needs to be done in this field. This is not very surprising, because relatively it is a very young research field that only arrose a little bit more than ten years ago. We are not yet close to the moment that affective human-machine interaction systems can be used in our everyday home environments, but there is certainly still a lot to be discovered in the years to come.

[1] Appeared in the episode Roboliefde *(Robolove)* of the VPRO programme De Toekomst *(The Future)*: http://www.vpro.nl/programma/detoekomst/afleveringen/26435176/ (in Dutch)

Acknowledgements

I would like to express a few words of appreciation because I am very grateful for the support, advise and friendship I could enjoy during the time I worked on this thesis.

First of all, I would like to thank my advisors, Tijn van der Zant, Fokie Cnossen (University of Groningen) and Christine Lisetti (Florida International University) for their guidance and for being a great source of ideas and inspiration. I thank Christine also for giving me the opportunity to work in her research group and providing the resources to do the project. Especially I owe Tijn my gratitude for making my project at FIU possible in the first place. Without him I would not even have considered my visit to Miami possible.

At FIU, I also received a lot of advise and great ideas from Armando Barreto at the Department of Electrical & Computer Engineering. I would like to thank him for providing the resources that I needed for doing my experiments and for his help with the use of the experimental set up.

Also I would like to send my thanks to the sweet friends I met in the United States. In particular, my stay in Miami would not have been the same without the friendship of two great co-workers, Francisco Ortega and Frank Hernandez and the coolest roommate I have ever had, Sandra Rios.

Of course I am also very grateful to my parents and Dutch friends for their constant care and support throughout the years and last but not least I would like to thank my boyfriend, Jaldert Rombouts, for being there all the time even though he was far away.

Chapter 7

Appendices

A-1 Emotion questionnaire

The following eight pages show the questionnaire that was given to the participants in the experiment for data collection. It includes the Ten Item Personality Inventory, the Emotion Wheel and the instructions that were given to the test subjects when they participated.

Emotion experiment

Written questionnaire accompanying the movie slideshow

Subject number: …….....

Welcome and thank you for participating in this experiment. We would like to ask you to first turn of your cell phone and fill out these three questions:

What is your age? ……...
What is your sex? ……...
What is your ethnicity? ……...

Now we would like to get to know you a little better by doing a very brief personality test. Please follow the instructions below.

Ten-Item Personality Inventory-(TIPI) (as described by Gosling et al. (2003))

Here are a number of personality traits that may or may not apply to you. Please write a number next to each statement to indicate the extent to which you agree or disagree with that statement. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other.

Disagree strongly 1	Disagree moderately 2	Disagree a little 3	Neither agree nor disagree 4	Agree a little 5	Agree moderately 6	Agree strongly 7

I see myself as:

1. _____ Extraverted, enthusiastic.
2. _____ Critical, quarrelsome.
3. _____ Dependable, self-disciplined.
4. _____ Anxious, easily upset.
5. _____ Open to new experiences, complex.
6. _____ Reserved, quiet.
7. _____ Sympathetic, warm.
8. _____ Disorganized, careless.
9. _____ Calm, emotionally stable.
10. _____ Conventional, uncreative.

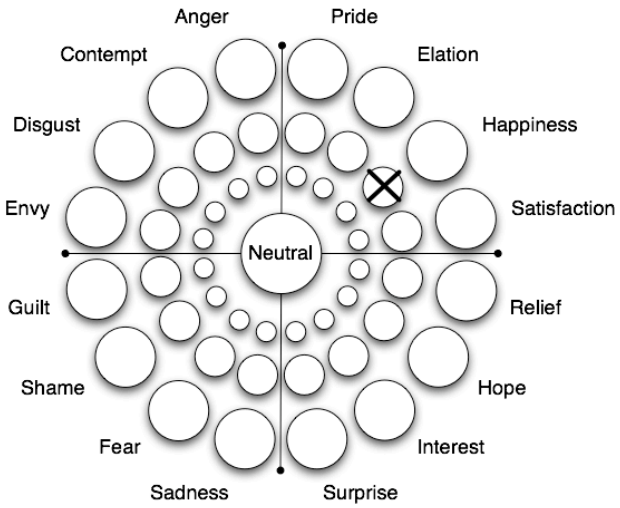
During this experiment you will watch a slideshow of movie fragments while your physiological signals are measured and your facial expressions are recorded on video. Try to keep your left arm, to which the sensors are attached as motionless as possible during the experiment, especially during the movies, and try to keep facing the screen (and camera) at all times during the movie fragments. Do not speak with the experimenter after the movie has started (unless something is wrong). After each movie fragment you will be asked to return to the written questionnaire and this is when we ask you to self-report your emotions. Indicate on the questionnaire how you felt when watching the movie fragment. Try to really indicate your own feelings, not the ones you project on the actors in the movies. We will guide you in this task by providing two questions:

Check the box for the emotion that you felt the most:
Example:

- ◊ Happy
- ◊ Angry
- ✓ Sad
- ◊ Disgust
- ◊ Surprise
- ◊ Fear
- ◊ Other, namely: ……

Indicate in the emotion wheel where you would place the emotion you felt while watching the movie:

Example:



Explanation: 16 different emotion families are arranged in a circular fashion. Please note that the word or label that represents each family can stand for a whole range of similar emotions. Thus, the Anger family also covers emotions such as rage, vexation, annoyance, indignation, fury, exasperation, or being cross or mad.
First identify approximately how you felt during the movie and choose the emotion family that best corresponds to the kind of feeling you experienced. Then determine with which intensity you experienced the respective emotion and check one of the circles in the "spike" corresponding to this emotion family -- the bigger the circle and the closer it is to the rim of the wheel, the stronger the emotional experience. If you felt no emotion at all, check the 'neutral' circle in the center.

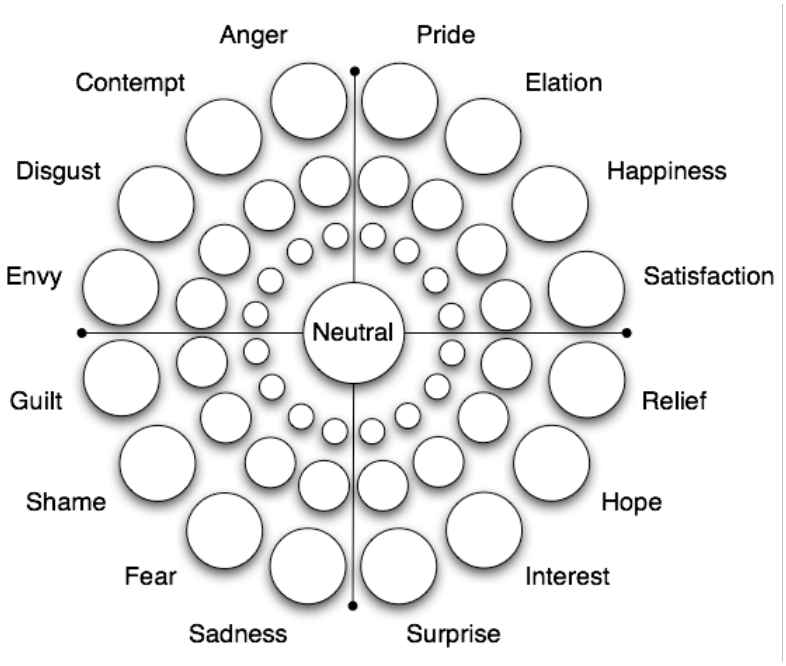
If you have any comments, you can write them down at the bottom. Now, we will start the slideshow! Follow the instructions on the screen and try not to move the sensors!

After watching Movie 1:

Check the box for the emotion that you felt the most during Movie 1:

- ◊ Happy
- ◊ Angry
- ◊ Sad
- ◊ Disgust
- ◊ Surprise
- ◊ Fear
- ◊ Other, namely: ……

Indicate in the emotion wheel where you would place the emotion you felt while watching Movie 1:



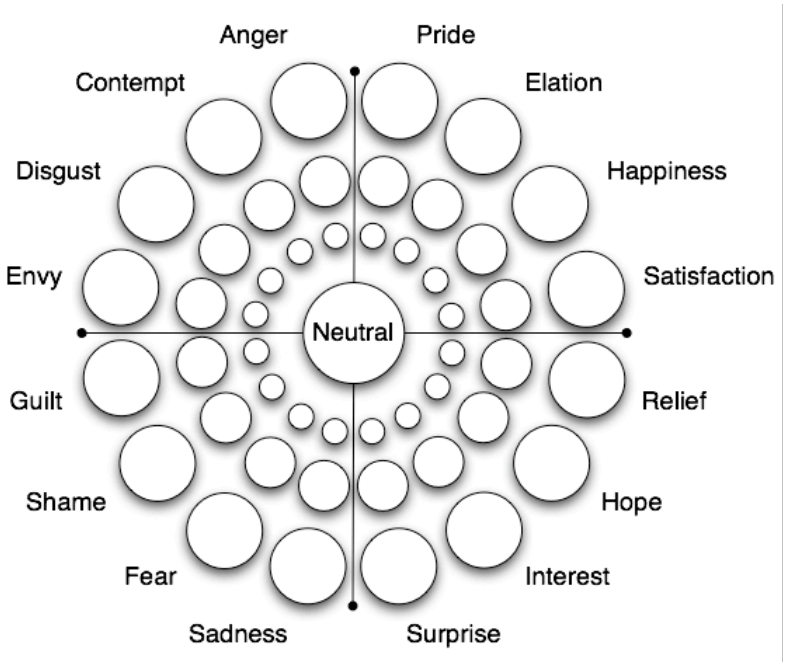
Comments:

After watching Movie 2:

Check the box for the emotion that you felt the most during Movie 2:

- ◊ Happy
- ◊ Angry
- ◊ Sad
- ◊ Disgust
- ◊ Surprise
- ◊ Fear
- ◊ Other, namely: ……

Indicate in the emotion wheel where you would place the emotion you felt while watching Movie 2:



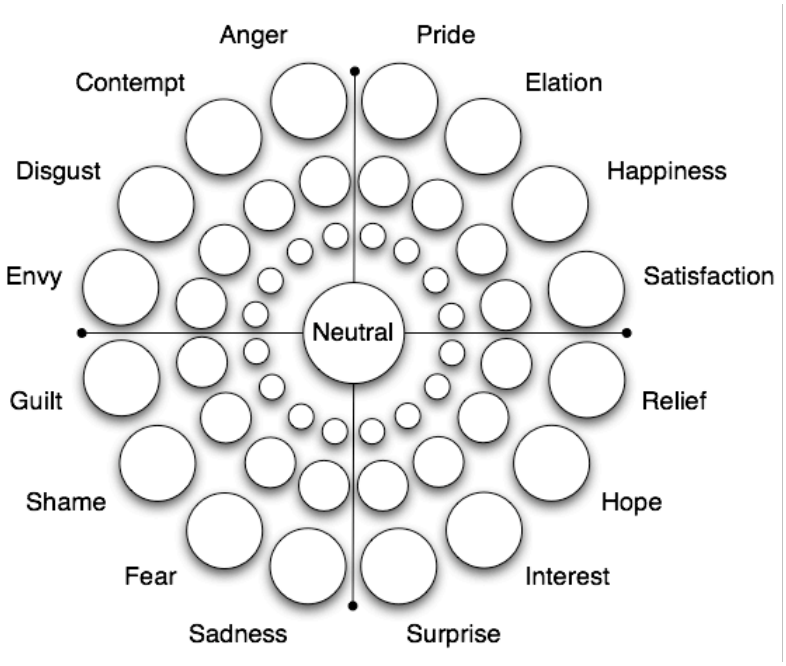
Comments:

After watching Movie 3:

Check the box for the emotion that you felt the most during Movie 3:

- ◊ Happy
- ◊ Angry
- ◊ Sad
- ◊ Disgust
- ◊ Surprise
- ◊ Fear
- ◊ Other, namely: ……

Indicate in the emotion wheel where you would place the emotion you felt while watching Movie 3:



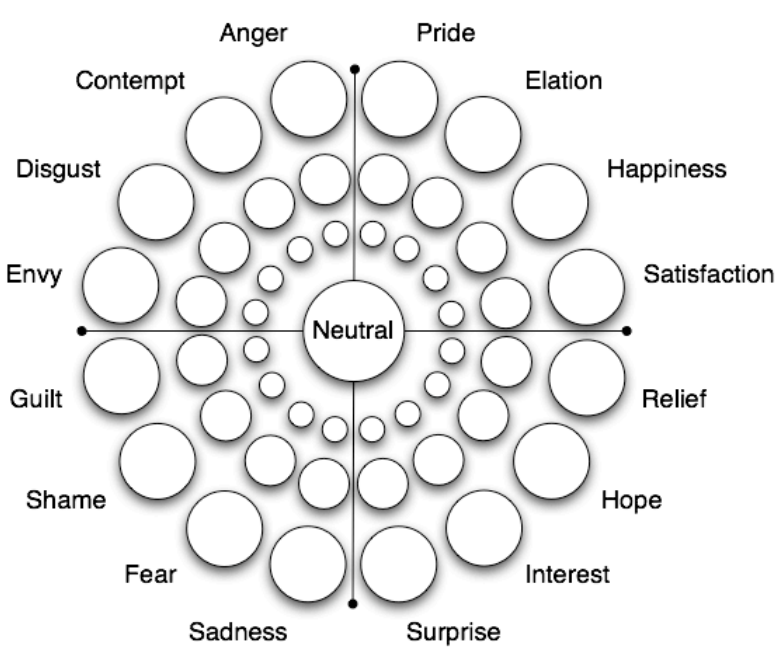
Comments:

After watching Movie 4:

Check the box for the emotion that you felt the most during Movie 4:

- ◊ Happy
- ◊ Angry
- ◊ Sad
- ◊ Disgust
- ◊ Surprise
- ◊ Fear
- ◊ Other, namely: ……

Indicate in the emotion wheel where you would place the emotion you felt while watching Movie 4:



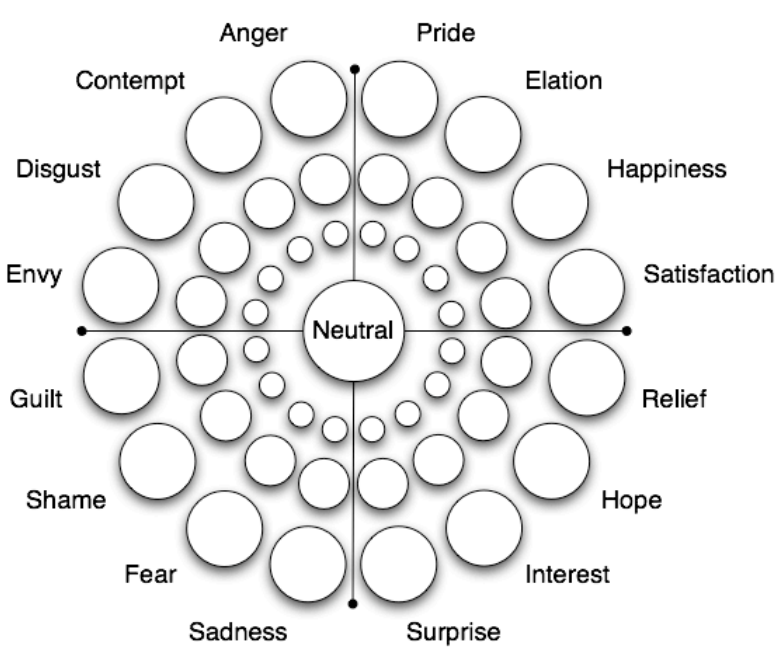
Comments:

After watching Movie 5:

Check the box for the emotion that you felt the most during Movie 5:

- ◊ Happy
- ◊ Angry
- ◊ Sad
- ◊ Disgust
- ◊ Surprise
- ◊ Fear
- ◊ Other, namely: ……

Indicate in the emotion wheel where you would place the emotion you felt while watching Movie 5:



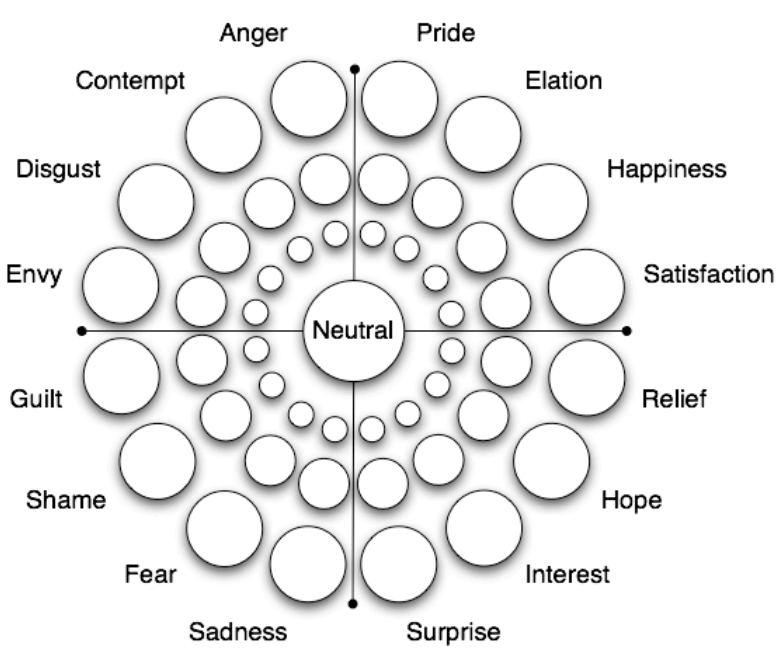
Comments:

After watching Movie 6:

Check the box for the emotion that you felt the most during Movie 6:

- ◊ Happy
- ◊ Angry
- ◊ Sad
- ◊ Disgust
- ◊ Surprise
- ◊ Fear
- ◊ Other, namely: ……

Indicate in the emotion wheel where you would place the emotion you felt while watching Movie 6:



Comments:

Bibliography

Arafa, Y. and Mamdani, A. (2003). Scripting embodied agents behaviour with cml: character markup language. In *IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, pages 313–316, New York, NY, USA. ACM.

Ark, W. S., Dryer, C. D., and Lu, D. J. (1999). The emotion mouse. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces-Volume I*, pages 818–823, Mahwah, NJ, USA. Lawrence Erlbaum Associates, Inc.

Ball, E. (2001). A bayesian heart: Computer recognition and simulation of emotion. *Emotions in Humans and Artifacts*, pages 303–332.

Barreto, A., Zhai, J., and Adjouadi, M. (2007). Non-intrusive Physiological Monitoring for Automated Stress Detection in Human-Computer Interaction. *Lecture Notes In Computer Science*, 4796.

Barrett, F. L. and Russell, J. A. (1998). Independence and bipolarity in the structure of affect. *Journal of Personality and Social Psychology*, 74(4):967–984.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J. (2005). Recognizing facial expression: machine learning and application to spontaneous behavior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 568–573 vol. 2.

Bartneck, C., Reichenbach, J., and Van Breemen, A. (2004). In Your Face, Robot! The Influence of a Characters Embodiment on How Users Perceive Its Emotional Expressions. *Proceedings of the Design and Emotion*.

Batliner, A., Fischer, K., Huber, R., Spilker, J., and Nöth, E. (2003). How to find trouble in communication. *Speech Commun.*, 40(1-2):117–143.

Bechara, A., Damasio, H., Tranel, D., and Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science*, 275(5304):1293–1295.

Biele, C. and Grabowska, A. (2006). Sex differences in perception of emotion intensity in dynamic and static facial expressions. *Experimental Brain Research*, 171(1):1–6.

Bower, G. H. (1981). Mood and memory. *Am Psychol*, 36(2):129–148.

Breazeal, C. and Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous Robots*, 12(1):83–104.

Breazeal, C. and Scassellati, B. (2000). Infant-like social interactions between a robot and a human caregiver. *Adaptive Behavior*, 8(1):49–74.

Bruce, A., Nourbakhsh, I., and Simmons, R. (2002). The role of expressiveness and attention in human-robot interaction. In *Robotics and Automation, 2002. Proceedings. ICRA '02. IEEE International Conference on*, volume 4, pages 4138–4142 vol.4.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *ICMI '04: Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211, New York, NY, USA. ACM.

Canamero, L. D. and Fredslund, J. (2000). How does it feel? emotional interaction with a humanoid lego robot. *Socially Intelligent Agents: The Human in the Loop. Papers from the AAAI 2000 Fall Symposium*, pages 23–28.

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Raouzaiou, A., and Karpouzis, K. (2006). Modeling naturalistic affective states via facial and vocal expressions recognition. *Proceedings of the 8th international conference on Multimodal interfaces*, pages 146–154.

Chow, C. and Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, 14(3):462–467.

Cohn, J. F. and Schmidt, K. (2003). The timing of facial motion in posed and spontaneous smiles. *Active Media Technology*.

Conati, C. (2002). Probabilistic assessment of user' s emotions during the interaction with educational games.

Conati, C., Gertner, A., and Vanlehn, K. (2002). Using bayesian networks to manage uncertainty in student modeling. *User Modeling and User-Adapted Interaction*, 12(4):371–417.

Cootes, T. F., Edwards, G. J., and Taylor, C. J. (1998). *Active Appearance Models*.

Cowie, R., Cowie, D. E., Savvidou, S., Mcmahon, E., Sawey, M., and Schröder, M. (2000). 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. ISCA.

Damasio, H., Grabowski, T., Frank, R., Galaburda, A. M., and Damasio, A. R. (1994). The return of Phineas Gage: clues about the brain from the skull of a famous patient. *Science*, 264(5162).

Demsar, J., Zupan, B., and Leban, G. (2004). Orange: From experimental machine learning to interactive data mining. Technical report, Faculty of Computer and Information Science, University of Ljubljana.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1):417–440.

Dishman, R. K., Nakamura, Y., Garcia, M. E., Thompson, R. W., Dunn, A. L., and Blair, S. N. (2000). Heart rate variability, trait anxiety, and perceived stress among physically fit men and women. *International Journal of Psychophysiology*, 37(2):121–133.

Dryer, D. C. (1999). Getting personal with computers: How to design personalities for agents. *Applied Artificial Intelligence*, pages 273–295.

Duda, R. O., Hart, P. E., and Stork, D. G. (1973). *Pattern classification and scene analysis*. Wiley New York.

Ekman, P. (1992). An argument for basic emotions. *Emotion: Themes in the Philosophy of the Mind*.

Ekman, P. and Friesen, W. V. (1978). *Facial action coding system*. Consulting Psychologists Press, Palo Alto, Calif.

Ekman, P. and O'Sullivan, M. (1991). Who can catch a liar? *The American psychologist*, 46(9):913–920.

Elliott, C., Rickel, J., and Lester, J. (1997). Integrating affective computing into animated tutoring agents. *Proceedings of the IJCAI Workshop on Animated Interface Agents: Making Them Intelligent*, 113121.

Esau, N., Kleinjohann, B., Kleinjohann, L., and Stichling, D. (2003). MEXI: machine with emotionally eXtended intelligence.

Fragopanagos, N. and Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405.

Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C., and Wang, J. (2005). Designing robots for long-term social interaction. In *Intelligent Robots and Systems, 2005. (IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 1338–1343.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

Grolleman, J., van Dijk, B., Nijholt, A., and van Emst, A. (2006). Break the habit! designing an e-therapy intervention using a virtual coach in aid of smoking cessation. pages 133–141.

Gross, J. J. and Levenson, R. W. (1995). Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108.

Hashimoto, T., Senda, M., Shiiba, T., and Kobayshi, H. (2004). Development of the interactive receptionist system by the face robot. In *SICE 2004 Annual Conference*, volume 2, pages 1404–1408 vol. 2.

Hilton, S. M. (1982). The defence-arousal system and its relevance for circulatory and respiratory control. *J Exp Biol*, 100(1):159–174.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Mach. Learn.*, 11(1):63–90.

Izard, C. E. (1971). *The face of emotion*. Appleton-Century-Crofts, New York.

Kapoor, A., Burleson, W., and Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human-Computer Studies*, 65(8):724–736.

Kapoor, A. and Picard, R. W. (2005). Multimodal affect recognition in learning environments. *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 677–682.

Kim, K., Bang, S., and Kim, S. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J. (1998). On combining classifiers. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(3):226–239.

Koenigs, M. and Tranel, D. (2007). Irrational Economic Decision-Making after Ventromedial Prefrontal Damage: Evidence from the Ultimatum Game. *Journal of Neuroscience*, 27(4).

Kraut, R. (2007). Aristotle's Ethics. *Stanford Encyclopedia of Philosophy*.

Kshirsagar, S. (2002). A multilayer personality model. *Proceedings of the 2nd international symposium on Smart graphics*, pages 107–115.

Kwon, O. W., Chan, K., Hao, J., and Lee, T. W. (2003). Emotion Recognition by Speech Signals. In *Eighth European Conference on Speech Communication and Technology*. ISCA.

Lang, P. J. (1995). The emotion probe. studies of motivation and attention. *The American psychologist*, 50(5):372–385.

Lang, P. J., Melamed, B. G., and Hart, J. (1970). A psychophysiological analysis of fear modification using an automated desensitization procedure. *Journal of abnormal psychology*, 76(2):220–234.

Lazarus, R. S. (1984). On the primacy of cognition. *American Psychologist*, 39(2):124–29.

Lee, C.-H. J., Kim, K., Breazeal, C., and Picard, R. (2008). Shybot: friend-stranger interaction for children living with autism. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3375–3380, New York, NY, USA. ACM.

Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, T. T., Stone, B. A., and Bhogal, R. S. (1997). The persona effect: affective impact of animated pedagogical agents. In *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 359–366, New York, NY, USA. ACM Press.

Leventhal, H. and Scherer, K. (1987). The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition and Emotion*, 1(1):3–28.

Lisetti, C. L., Brown, S., Alvarez, K., and Marpaung, A. (2004). A social informatics approach to human-robot interaction with an office service robot. *IEEE Transactions on Systems, Man, and Cybernetics–Special Issue on Human Robot Interaction*, 34(2):195–209.

Lisetti, C. L. and Maurpang, A. (2006). BDI+ E Framework: An affective cognitive modeling for autonomous agents based on scherer's emotion theory. *Proceedings of KI06 29th German Annual Conference in Artificial Intelligence, Bremen, Germany*.

Lisetti, C. L. and Nasoz, F. (2002). MAUI: a multimodal affective user interface. *Proceedings of the tenth ACM international conference on Multimedia*, pages 161–170.

Maat, L. and Pantic, M. (2007). Gaze-x: Adaptive, affective, multimodal interface for single-user office scenarios. pages 251–271.

Mehrabian, A. (2007). *Nonverbal Communication*. Aldine.

Melder, W. A., Truong, K. P., Den Uyl, M., Van Leeuwen, D. A., Neerincx, M. A., Loos, L. R., and Plum, S. B. (2007). Affective multimodal mirror: sensing and eliciting laughter. In *HCM '07: Proceedings of the international workshop on Human-centered multimedia*, pages 31–40, New York, NY, USA. ACM.

Merkx, P. A. B., Truong, K. P., and Neerincx, M. A. (2007). Inducing and measuring emotion through a multiplayer first-person shooter computer game. *Proceedings of the Computer Games Workshop*.

Minato, T., Shimada, M., Ishiguro, H., and Itakura, S. (2004). Development of an android robot for studying human-robot interaction. pages 424–434.

Mori, M. (1970). The Uncanny Valley. *Energy*, 7(4):33–35.

Murphy, K. (2001). The bayes net toolbox for matlab. *Computing Science and Statistics*, 33(2):1024–1034.

Nasoz, F., Alvarez, K., Lisetti, C. L., and Finkelstein, N. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1):4–14.

Nasoz, F. and Lisetti, C. (2007). Affective user modeling for adaptive intelligent user interfaces. pages 421–430.

Nasoz, F. and Lisetti, C. L. (2006). MAUI avatars: Mirroring the user's sensed emotions via expressive multi-ethnic facial avatars. *Journal of Visual Languages and Computing*, 17(5):430–444.

Niedenthal, P. M. (2007). Embodying emotion. *Science*, 316(5827):1002–1005.

Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

Ortony, A. and Turner, T. J. (1990). What's basic about basic emotions? *Psychol Rev*, 97(3):315–331.

Ostermann, J. (1998). Animation of synthetic faces in mpeg-4. In *CA '98: Proceedings of the Computer Animation*, Washington, DC, USA. IEEE Computer Society.

Paleari, M. and Lisetti, C. (2006a). Psychologically grounded avatars expressions. In *29th Annual Conference on Artificial Intelligence*.

Paleari, M. and Lisetti, C. L. (2006b). Toward multimodal fusion of affective cues. *Proceedings of the 1st ACM international workshop on Human-centered multimedia*, pages 99–108.

Panksepp, J. (1982). Toward a general psychobiological theory of emotions. *Behavioral and Brain Sciences*, 5(3):407–467.

Pantic, M. and Rothkrantz, L. J. M. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390.

Pantic, M., Sebe, N., Cohn, J. F., and Huang, T. (2005). Affective multimodal human-computer interaction. *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 669–676.

Pavlovic, V., Garg, A., Rehg, J. M., and Huang, T. S. (2000). Multimodal Speaker Detection using Error Feedback Dynamic Bayesian Networks. *Proc. of Conf. on Computer Vision and Pattern Recognition*, 51:34–41.

Peter, C. and Herbon, A. (2006). Emotion representation and physiology assignments in digital systems. *Interacting with Computers*, 18(2):139–170.

Picard, R. W. (1995). Affective computing. Technical report, MIT.

Picard, R. W. (1997). *Affective computing*. MIT Press, Cambridge, MA, USA.

Picard, R. W. (1999). Affective Computing for HCI. *Human-Computer Interaction: Ergonomics and User Interfaces*, 1:829–833.

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

Prendinger, H., Mori, J., Saeyor, S., Mori, K., Okazaki, N., Juli, Y., Mayer, S., Dohi, H., and Ishizuka, M. (2004). Scripting and evaluating affective interactions with embodied conversational agents. *Künstliche Intelligenz (KI) Zeitschrift*, 1:4–10.

Puri, C., Olson, L., Pavlidis, I., Levine, J., and Starren, J. (2005). StressCam: non-contact measurement of users' emotional states through thermal imaging. *Conference on Human Factors in Computing Systems*, pages 1725–1728.

Rosis, F., Novielli, N., Carofiglio, V., Cavalluzzi, A., and Carolis, B. D. (2006). User modeling and adaptation in health promotion dialogs with an animated character. *Journal of Biomedical Informatics*, 39(5):514–531.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Russell, J. A. and Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805–819.

Russell, J. A. and Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273–294.

Russell, S. J. and Norvig, P. (1995). *Artificial intelligence: a modern approach*, chapter 21. Prentice-Hall, Inc. Upper Saddle River, NJ, USA.

Scherer, K. R. (2001). Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, pages 92–120.

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729.

Scherer, K. R. and Ellgring, H. (2007). Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion (Washington, D.C.)*, 7(1):113–130.

Sebe, N., Cohen, I., Gevers, T., and Huang, T. S. (2006). Emotion recognition based on joint visual and audio cues. In *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, pages 1136–1139, Washington, DC, USA. IEEE Computer Society.

Sharma, R., Pavlovic, V. I., and Huang, T. S. (1998). Toward multimodal human-computer interface. *Proceedings of the IEEE*, 86(5):853–869.

Sosnowski, S., Kuhnlenz, K., and Buss, M. (2006). Eddie - an emotion-display with dynamic intuitive expressions. In *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, pages 569–574.

Tian, Y. I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(2):97–115.

Tomkins, S. S. (1962). Affect, imagery, consciousness (Vol. 1). *New York*.

Valstar, M. F., Pantic, M., Ambadar, Z., and Cohn, J. F. (2006). Spontaneous vs. posed facial behavior: automatic analysis of brow actions. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 162–170, New York, NY, USA. ACM.

van Breemen, A. J. N. (2004). Animation engine for believable interactive user-interface robots. In *Intelligent Robots and Systems, 2004. (IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on*, volume 3, pages 2873–2878 vol.3.

van Kuilenburg, H., Wiering, M., and den Uyl, M. (2005). A model based method for automatic facial expression recognition. pages 194–205.

Wagner, J., Kim, J., and Andre, E. (2005). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 940–943.

Walker, J. H., Sproull, L., and Subramani, R. (1994). Using a human face in an interface. In *CHI '94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 85–91, New York, NY, USA. ACM Press.

Watson, D. and Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219–235.

Wierzbicka, A. (1992). Defining emotion concepts. *Cognitive Science*, 16(4):539–581.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. 2 edition.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist*, 35(2):151–175.

Zeng, Z., Fu, Y., Roisman, G. I., Wen, Z., Hu, Y., and Huang, T. S. (2006). Spontaneous Emotional Facial Expression Detection. *Journal of Multimedia*, 1(5):1–8.

Zeng, Z., Tu, J., Liu, M., Huang, T. S., Pianfetti, B., Roth, D., and Levinson, S. (2007). Audio-visual affect recognition. *Multimedia, IEEE Transactions on*, 9(2):424–428.

Zhang, T., Johnson, H. M., and Levinson, S. E. (2004). Childrens Emotion Recognition in an Intelligent Tutoring Scenario. *Proc. of ICSLP*.