Machine beats human emotion recognition through audio, visual, and physiological modalities

Joris H. Janssen

Eindhoven University of Technology / Philips Research

Paul Tacken, Gert-Jan de Vries

Philips Research

Egon L. van den Broek

University of Twente / Radboud University Medical Center Nijmegen

Joyce H.D.M. Westerink

Philips Research

Wijnand A. IJsselsteijn

Eindhoven University of Technology

Pim Haselager

Radboud University Nijmgen

Joris H. Janssen
Human Technology Interaction, Dep. of IE&IS
Eindhoven University of Technology
Den Dolech 2, 5600 MB Eindhoven, The Netherlands
e-mail: joris.h.janssen@philips.com
tel: +31 (0)6 1782 0353

Abstract

Over the last decade, an increasing number of studies have focused on automated recognition of human emotions by machines. However, performances of machine emotion recognition studies are difficult to interpret because benchmarks are lacking. In order to provide such a benchmark, we compared machine with human emotion recognition. Facial expressions, speech, and physiological signals were gathered from 17 individuals expressing 5 different emotional states. Support Vector Machines achieved a 82% recognition accuracy based on a physiological and facial features. In experiments with 75 humans on the same data, a maximum recognition accuracy of 57% was obtained. As machines outperform humans, automated emotion recognition is likely to be ready to be put into practice, which enables many new applications of human computer interaction.

Keywords: Affective computing, Emotion, Machine learning, Psychophysiology, Speech, Facial expressions

Introduction

Although science has been interested in emotions for a long time (Darwin, 1872; James, 1884), only relatively recently have the engineering and computer sciences developed an interest in emotion research. This increased interest is often attributed to Picard's book Affective Computing (Picard, 1997), in which she baptizes a field that studies artificial systems that are able to detect and recognize user's emotions, can appropriately express emotions, and may even feel emotions. As a truly interdisciplinary field, the challenges Affective Computing faces require the integration of work from many different disciplines (Calvo & D'Mello, 2010; Kappas, 2010). Hence, since the launch of Picard's book, there has been work by computer scientists on developing emotion algorithms (Zeng, Pantic, Roisman, & Huang, 2009), by engineers on developing (wearable) sensors (Westerink et al., 2009), and by psychologists on improving methodologies to take their research out of the lab and into the real life (Wilhelm & Grossman, 2010).

All these efforts are warranted by the seemingly endless possibilities for innovations and applications of affective technology. From a traditional human-computer interaction (HCI) perspective, affective computing examples involve, for instance, computers detecting and adapting to user frustration (Scheirer, Fernandez, Klein, & Picard, 2002) to make computer use more pleasant and satisfying (Whang, 2008). Other domains that can benefit from affective technologies include, for instance, cars that could track the driver's emotional state. This could improve safety by reducing music volume, activating the cell phone silencer, or even take over driving functionality in difficult situations (Healey & Picard, 2005; Katsis, Katertsidis, Ganiatsas, & Fotiadis, 2008). Tutoring systems could adapt to a student's affective state to make education more effective (Sarrafzadeh, Alexander, Dadgostar, Fan, & Bigdeli, 2008). Communication between humans could be improved by adding affective channels, like physiological signals, which can make the communication more intimate (Janssen, Bailenson, IJsselsteijn, & Westerink, 2010). In another vein, relaxation systems using affective computing can enable personalized strategies to help people cope with stress. In this light, music could be selected to relax or energize someone (Janssen, Van den Broek, & Westerink, 2011) and movies could be enhanced by using, for instance, affective haptic feedback (Lemmens, Crompvoets, Brokken, Van den Eerenbeemd, & De Vries, 2009). This is only a tip of the iceberg of innovations based on affective computing technology that may be awaiting us.

For all the proposed applications, some form of machine emotion recognition is necessary. This has not gone unnoticed during the last decade, with machine emotion recognition often being treated as the holy grail of affective computing (Cowie, 2009; Picard, 2003; Picard & Klein, 2002). A large number of studies have tried to recognize human emotions from facial expressions, posture, speech, and physiological indices (see the Background section for a review). Despite all these studies and several extensive surveys (Cowie et al., 2001; Van den Broek et al., 2010; Zeng et al., 2009), many researchers seem to agree that automated emotion recognition is not yet mature enough to be put into practice (Calvo & D'Mello, 2010; Kappas, 2010). This may be due to the lack of clarity concerning what kind or level of performance is necessary for successful applications. In fact, it is even unclear what the general standards of emotion recognition performance are, which makes it problematic to assess the current state-of-the-art performance.

We challenge the claim that the current state-of-the-art of machine emotion recognition is not good enough for practical application. Instead, we would like to pose that we do not know if machine emotion is currently developed far enough to be put into practice. First of all, we do not know what kind of performance is necessary for successful affective computing applications. The precision and accuracy with which an affective state has to be identified depends on the specific system or product in which the affective technology is applied. For a music player as mentioned above, distinguishing between relaxed and energetic (which could be relatively easy to do) can already be useful. However, simply distinguishing relaxed and energetic states might not be sufficient for, for instance, affective technology that can train people in their social interaction. Hence, as the necessary performance depends on the application, there might well be applications for which the current state-of-the-art is sufficient.

Second, it is problematic to interpret performance of different machine emotion recognition approaches because there is a lack of real-world testing. Machine emotion recognition that is developed is often tested solely in the lab (e.g., J. Kim & André, 2008; Lisetti & Nasoz, 2004; Zhai & Barreto, 2006). Therefore, it is unclear how these systems carry over to the real world. Real world validation is necessary to see if the system can deal with complexities apparent outside the lab.

Third, it is difficult to compare different machine emotion recognition approaches because there is a lack of benchmarks. Benchmarks could be obtained by either testing different machine emotion recognition approaches in (the same) practical applications, or by comparing them against a common standard in emotion recognition. However, most machine emotion recognition approaches test systems that are disconnected from actual applications (e.g., Picard, Vyzas, & Healey, 2001; Rani, Liu, Sarkar, & Vanman, 2006; Van den Broek et al., 2010). In other words, only the emotion recognition component is tested, instead of an entire application in which the emotion recognition component is supposed to be embedded in. This is useful when developing generic emotion recognition components that can be used in different applications. Furthermore, there are no common standards to compare machine emotion recognition against. Therefore, it is very difficult, if not impossible, to compare different machine emotion recognition approaches.

We propose to use human performance as a starting point for benchmarking machine emotion recognition. First of all, humans can recognize emotions and report about what they recognized in an understandable way. Second, other applications of artificial intelligence (including machine learning) are also benchmarked against human performance; for instance, playing chess (F.-H. Hsu, 2002) and recognizing characters (Chellapilla, Larson, Simard, & Czerwinski, 2005). Third, humans are generally recognized as the most superior natural systems performing mental tasks (Roth & Dicke, 2005). Moreover, traditionally artificially intelligent applications are seen as successful if they are able to beat human performance Turing (1950). A recent example of this is Watson, a machine that was able to beat expert humans in a question and answer game show (Ferucci et al., 2010). Hence, human performance seems like a good starting point as a benchmark for machine emotion recognition. Moreover, it allows for comparison of different modalities, as humans use both visual and aural modalities to recognize emotions (Zaki, Weber, Bolger, & Ochsner, 2009).

In the current work, we set out to investigate how well machine emotion recognition performs by comparing it to human emotion recognition. Furthermore, we investigate what the relative performances are of emotion recognition based on (combinations of) different sources of emotional information (i.e., modalities). This way, we hope to quantify where the current state-of-the-art of machine emotion recognition stands and what fruitful directions for future research and technologies might be.

The rest of this manuscript is organized as follows. In the background section, we first review the main theories on emotions and how emotions could be measured and automatically recognized. Subsequently, we perform a meta-analysis of studies that have investigated machine emotion. Following the background section, we describe how we gathered the data that we used for our machine learning algorithms and our human benchmark. Thereafter, the human benchmark and its results are presented. This is followed by the machine learning section, in which we describe our feature extraction and classification procedures together with the results, and compare them to the human benchmark. Finally, the findings are discussed and limitations and future directions are given.

Background

In this Background section, we will first explain what emotions are, what bodily processes are involved in emotions, and how these bodily signals can be measured. It is not our intention to provide a complete review of the entire field, but to just sketch some of the issues involved that are most relevant for our work. The interested reader is referred to Lewis, Haviland-Jones, and Barrett (2008), Petta, Pelachaud, and Cowie (2011), and K. R. Scherer, Bänziger, and Roesch (2010) for a more complete overview of emotion and affective computing research. In the second part of this section, we perform a survey of machine learning studies that try to recognize emotions, and draw conclusions on the results so far.

Emotions and expressions of emotions

The debate on emotion has largely been characterized by two opposing theories: the James-Lange theory and the Cannon-Bard theory. On the one hand, the James-Lange theory focuses on the bottom-up processes involved in emotion generation, by claiming that "our feeling of the [bodily] changes as they occur is the emotion" (James, 1884, p. 190). Later theories inspired by the James-Lange theory include behavioral theory (Ryle, 1949) , the somatic marker hypothesis by Damasio (1994), perceptual theory of emotions by Prinz

(2004) , and the work of Zajonc (1980) and Frijda (1986). Although these theories differ on some aspects, all argue for a bottom-up approach to emotion generation, starting from bodily states. On the other hand, theories inspired by Cannon-Bard Cannon (1927) argue for more top-down (or cognitive) influences in emotion generation. Here, dimensional appraisal theories like the theories of Arnold (1960) and Lazarus (1991) have received a lot of attention. According to dimensional appraisal theories, emotions include appraisal judgments (i.e., judgments to the effect that one is facing a situation or predicament that matters). These theories treat bodily states as the result of appraisal judgments and therefore take a more top-down view on emotion generation.

Recently, the top-down view and the bottom-up view have started to become integrated into a synthetic perspective acknowledging that both processes play a role in emotion generation (K. R. Scherer & Zentner, 2001). In line with this, neuroscientists have identified pathways for bottom-up processes (LeDoux, 2000; Phelps, 2006) and top-down processes (Phelps et al., 2001; Teasdale et al., 1999). In a recent paper, Ochsner et al. (2009) explicitly compared bottom-up and top-down processes in emotion generation. They found evidence for a dual pathway model in which bottom-up processes primarily activate the amygdala and top-down processes activate prefrontal regions that represent high level cognitive interpretations. This further confirms that bodily expressions and cognitive processes both play a significant role in emotion generation.

For our current purposes, it is important to recognize that emotions are strongly coupled to facial expressions, posture, speech prosody, and physiological responses. It does not matter whether this is a top-down or bottom-up coupling, or a hybrid form as is currently the main model. The key point is that we can find correlates of emotions in these bodily responses, as all theories acknowledge. Facial expressions are probably the most widely studied (Ekman, 1992; Izard, 1971). Findings from this research show that specific emotions relate to specific facial expressions (Ekman, 1992, 1999). Moreover, many of these expressions seem to be universally similar (Ekman, 2009; Ekman & Friesen, 1971, 1986; Matsumoto & Willingham, 2009). In addition, facial expressions are thought to mostly signal the pleasantness of an emotion as opposed to the arousal of the emotion. More recently, posture has been studied as a strong way of expressing and communicating emotions (Hoogen, IJsselsteijn, & Kort, 2008; Gelder, 2009; Gelder et al., 2010; Kleinsmith, Bianchi-Berthouze, & Steed, 2011). Besides muscular emotion expressions, emotions are also related to the activity of the sympathetic and parasympathetic nervous systems. These systems connect to our internal organs and their activity can be inferred from, for instance, our heart rhythm, respiration, skin conductance and temperate, and blood volume pulse. Ekman, Levenson, and Friesen (1983) were one of the first who showed that different emotions relate to different physiological patterns, and they were followed by many others (see Cacioppo, Berntson, Klein, & Poehlmann, 1997, and Cacioppo, Tassinary, & Berntson, 2000, for meta-analyses). Physiological signals are thought to signal mostly the arousal of the emotion as opposed to the pleasantness. Finally, emotions are also expressed through speech prosody and nonlinguistic vocalizations (Bachorowski & Owren, 2008; Banse & Scherer, 1996; K. R. Scherer, 1986). Taken together, emotions have many measurable aspects, including physiological, facial, postural, and vocal signals.

Automated emotion recognition can be done based on (a combination of) these different signals. When considering the requirements on sensors, there are different advantages and disadvantages to each method. Facial expression and posture recognition requires the user to be in front of a camera. Applications that involve the user being in a static position (e.g., sitting behind a desk, in a car) are most suited for this. When the user is moving, it is considerably more difficult to track facial expression and posture. One approach could be to use EMG electrodes positioned on the face or body (Picard, 2000), but this can be obtrusive. Speech is especially useful for tracking conversations and can easily be measured unobtrusively (Pentland, 2005, 2008). In many other settings, speech is absent and therefore unsuitable for emotion tracking. Finally, it is possible to measure noninvasive physiological signals in a mobile context, but the sensors used for this are sometimes considered obtrusive. Fortunately, over the last decade there have been many advances to wearable physiological measurement platforms, and sensors are being integrated into all kinds of devices (Westerink et al., 2009). Hence, physiological sensors are receiving an increasing interest in affective computing applications.

A completely different way of assessing emotions is by asking people how they feel. Subjective feeling is often considered as one of many aspects of emotions (J. A. Russell, 2003). Subjective feelings can be measured using self-report methods. Although there is a rich set of self-report tools that assess affective states, many of these are based on the finding that self-reported emotions can largely be accounted for in two dimensions (Lang, 1995; Mehrabian, 1970). The most common dimensions are valence (or pleasantness) and arousal. The dimensional approach to emotion self-report is a popular method as it provides a way of quantifying the emotional state and requires only two questions which can be answered relatively quickly. Furthermore, its validity has been extensively studied (Bradley, Codispoti, Cuthbert, & Lang, 2001; Bradley & Lang, 1994; Lang, Bradley, & Cuthbert, 1998). Nonetheless, asking people how they feel is far from straightforward because of problems relating to the reliability of introspection (e.g., Lyons, 1986), and because of the factors influencing subjects' verbalizations (e.g. privacy, the social (un)desirability of particular answers). Moreover, processing the answers is not possible in many automated emotion recognition applications. However, self-report data can serve as a ground truth for statistical learning methods that try to predict the self-reported emotion from automated measurements.

A survey of machine learning studies on affect recognition

Machine learning of affect is essentially a pattern recognition problem. The goal of pattern recognition is to develop an artificial system that is able to recognize (complex) patterns, in our case emotions, through statistical learning techniques. It follows the typical pattern recognition processing pipeline (see also Figure 1 and Meisel, 1972: a signal that is captured and, subsequently, processed by a physical system (e.g., a webcam, a PC's audio card, or a physiological recording device). This system provides us with the raw signals (e.g., an image, audio track, or physiological signal) on which preprocessing (e.g., noise reduction) and/or feature extraction (and selection) is applied. This results in a pattern space. The pattern space, which incorporates the selected features, is used for the pattern classification process. This classification process can either be the development of the classifying system or the execution of an already developed classifier on a new set of data. In the former case, the decision rule for the classifier is developed; in the latter case, the classification process provides a label for the signal that was captured. Classifying

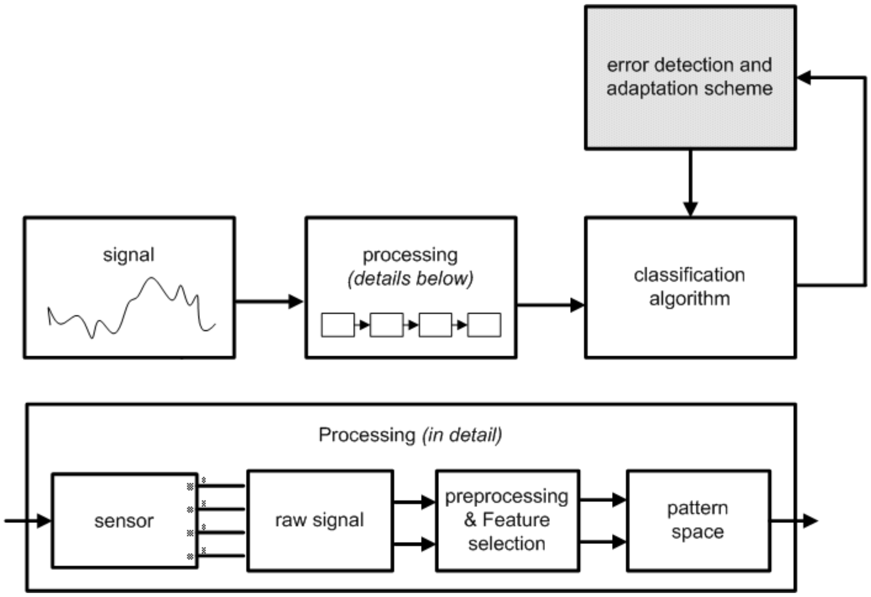
error detection and adaptation scheme
signal
processing (details below)
classification algorithm
Processing (in detail)
sensor
raw signal
preprocessing & Feature selection
pattern space

Figure 1. The pattern recognition processing pipeline, inspired by the work of Meissel (1972). Signals are collected from sensors and subsequently processed resulting in pattern space. This pattern space forms the input for the classification algorithm. The gray box is utilized with supervised classification, as it determines the error and the adaptation of the classification process. With unsupervised learning the decision algorithm is fixed, as no a priori knowledge is available on which the error detection can be based.

systems can apply either template matching, syntactic or structural matching, or statistical classification (e.g., artificial neural networks). In affective computing, the former two are not or seldom used. Regarding the third, if a label or category to which the measurement space belongs is available, a classification error can be determined and the classification process can be adapted (i.e., supervised learning). This labeled set of data is often denoted as the training data. Statistical pattern recognition uses input features, a discriminant function (or network function for artificial neural networks) that takes the input features to recognize the classes, and an error criterion in its classification process. Figure 1 visualizes this machine learning pipeline.

Tables 2-5 review 48 different studies that have employed pattern recognition techniques to classify different emotional states from facial/video, speech/audio, and/or physiological signals. The machine learning pipeline can be employed for each data source (i.e., modality) separately or, when the features from all data sources are extracted, they can be merged into one set. Both approaches are applied frequently. Tables 2-5 illustrate both differences and similarities in research on affective computing that utilized distinct modalities. All studies comply with the typical pattern recognition processing pipeline. However, many differences can be identified between and within the tables. In the following paragraphs, we will discuss the most important common characteristics of studies depicted in the tables.

To enable processing of the signals that are expected to reflect emotions, in most cases comprehensive sets of features have to be identified for each affective signal. To extract these features, the affective signals are processed in the time (e.g., statistical moments), frequency (e.g., Fourier), time-frequency (e.g., wavelets), or power domain (e.g., periodogram and autoregression). In Tables 2-5 it is indicated what signals were recorded in the reported studies and what features were derived from them. When inspecting Tables 2-5, it is remarkable that the number of features extracted is very different for the different studies. Scheirer et al. (2002) report the use of only five features and (Van den Broek, 2011) used only four features. In contrast, Littlewort, Bartlett, Fasel, Susskind, and Movellan (2006) report 900 features, Chanel, Kierkels, Soleymani, and Pun (2009) report using more than 2000 features, and Xiao, Zhao, Zhang, and Shi (2011) even report using 4320 features. Only half of the studies applied feature selection/reduction, where this would be advisable in general. Feature selection/reduction identifies the most important features and discards features that merely generate noise for the classification systems. The need for feature selection/reduction is lower when a small set of features is explored.

Not only the number of features differs considerably between studies, but also the number of participants ranges from 1 (Busso et al., 2004; Picard et al., 2001) up to 100 or even more (K. H. Kim, Bang, & Kim, 2004; Lien, Kanade, Cohn, & Li, 2000; Littlewort et al., 2006; Xiao et al., 2011). This number of participants used is an important parameter, as it gives information as to how reliable the findings our when applied to a larger population. Studies including more than 30 participants are relatively rare, which makes their results difficult to generalize to other populations.

For affective computing, a plethora of classifiers is used, as is illustrated in Tables 2-5. Classification in affective computing is hard as the emotion classes used are typically ill defined. This makes it difficult to compare studies. Moreover, the number of emotion categories to be discriminated in the different studies ranges considerably: from 2 to 4 in most studies up to around twelve (Banse & Scherer, 1996; Gunes & Piccardi, 2009). These are relatively small numbers of classes in terms of pattern recognition and machine learning (e.g., compare this with 26 letters plus ten digits in handwriting recognition). It is notable that the recognition rates of affective classifications are typically lower than recognition rates in other classification problems. In affective computing, correct recognition rates of 60% – 80% are common; see Tables 2-5. In contrast, in most other pattern recognition problems, recognition rates of over 90% (and often over 95%) are achieved (Jain, Duin, & Mao, 2000); for example, multimedia analysis (Jung, Kim, & Jain, 2004), optical character recognition (OCR; Mori, Suen, & Yamamoto, 1992), handwriting recognition (Lorigo & Govindaraju, 2006), and face recognition (Bowyer, Chang, & Flynn, 2004). This illustrates the complex nature of affective computing.

To further review previous work on affective computing we discuss three modalities that have frequently been employed for machine emotion recognition: vision/image, audio/speech, and physiological signals. These three modalities can be used to analyze facial expressions, speech utterances, and physiological processes. In the next sections we will first discuss each of these three modalities separately. Subsequently, we will discuss work on the combination of these three modalities.

Vision-based emotion recognition. Vision-based emotion recognition is mostly done based on facial expression analysis. Facial expression analysis is complex, at least in part because physiognomies of faces vary considerably among individuals due to age, ethnicity, gender, facial hair, cosmetic products, and occluding objects (e.g., glasses and hair). Furthermore, a face can appear to be distinct from itself due to pose and/or lighting changes, or other forms of environmental noise. For a more elaborate discussion on these issues, we refer to the surveys of Fasel and Luettin (2003) and Tian, Kanade, and Cohn (2005).

Despite its complexity, emotion recognition based on facial expression remains popular within the scientific community. This is likely because, it is among human's dominant modalities for emotion recognition, and it can be applied to stimuli of different types (e.g., still images and video). Second, Ekman and Friesen's FACS (Ekman & Friesen, 1978) pro-

Table 1: Explanations for the abbreviations used in Tables 2-5.

Abbreviation	Explanation
AdaBoost	Adaptive Boosting
ANFIS	Adaptive-Network-based Fuzzy Inference Systems
ANOVA	Analysis of Variance
BN	Bayesian Network
C	Cardiovascular activity (e.g., ECG and BVP)
CTKF	Continuous Time Kalman Filter
DT	Decision Ttree
E	Electrodermal activity
EMDC	A tailored ensemble of binary classifiers
FACS	Facial Action Coding System markers
Feats	Number of features
GA	Genetic Algorithm
GMM	Gaussian Mixture Models
GP	Gaussian Process classification
GWF	Gabor Wavelet Filters
HMM	Hidden Markov Models
K*	Instance-based classifier, with a distance measure based on entropy
LDA	Fisher's Linear Discriminant Analysis
LRM	Linear Regression Model
LPP	Locality Preserving Projection
kNN	k-Nearest Neighbors
Mo	Movement
M	Electromyogram
NB	Nave-Bayes
PCA	Principal Component Analysis
Pr	Pressure
R	Respiration
RT	Regression Tree
Perf	Performance
SBS	Sequential Backward Selection
SFS	Sequential Forward Selection
Sp	Speech
Ss	Subjects
SUD	Subjective Unit of Distress
SVM	Support Vector Machines
T	Skin temperature
Vi	Vision
VFI	Voted Features Interval

Table 2: Review of twelve machine learning studies employing computer vision to recognize emotions.

References	Signals	Ss	Feats	Functions	Targets	Perf
Cottrell & Metcalfe (1991)	Face	20	4096	NN	8 emotions	20%
Essa & Pentland (1995; 1997)	FACS	8			5 emotions	98%
Yacoob & Davis (1996)	Mouth	32	16		7 emotions	65%
Lien et al. (2000)	FACS	100	38	LDA, HMM	9 action units	80%
Cohen et al. (2003)	Motion Units	53	12	BN	7 emotions	83%
Zhang & Ji (2005)	FACS		24	BN	6 emotions	72%
Pantic & Patras (2006)	FACS	19	24		30 action units	87%
Littlewort et al. (2006)	Face	100	900	LDA, SVM	7 emotions	93%
Gunes & Piccardi (2007)	Head	4	148	BN	6 emotions	75%
	Body		140			90%
	Head, body		288			94%
Gunes & Piccardi (2009)	Head, hands, body	10	172	DT, BN, SVM, ANN, AdaBoost	12 emotions	85%
Sanchez et al. (2010)	Face	52	84	SVM	6 emotions	95%
Xiao et al. (2011)	Face	210	4320	ANN, SVM	6 emotions	97%

Notes. Abbreviations are explained in Table 1.

vide an excellent set of markers that can be employed for vision-based emotion recognition. This provides a rather solid theoretical foundation and, consequently, FACS is often used as a basic starting point. Third, vision-based facial analysis is very well possible in controlled conditions. Par excellence, this is illustrated by the inclusion of face recognition in the biometrics portfolio (Van den Broek et al., 2010). Most affective computing research is still conducted within a controlled environment (e.g., a laboratory) and, hence, vision-based emotion recognition can be applied conveniently.

Xiao et al. (2011) showed that thousands of features can be derived from human faces. These features can be characterized along several dimensions (Fasel & Luettin, 2003). First, local features versus the face as a whole can be distinguished. In both cases segmentation of the image is required, often augmented by a priori knowledge of human observers. Second, deformation analysis versus motion extraction can be distinguished. With deformation analysis, a neutral image or a face model is compared with the image of interest. Motion extraction directly compares image sequences with each other.

As is shown in Table 2, in practice, the number of features extracted varies considerably: from 16 (Yacoob & Davis, 2006) to 4320 (Xiao et al., 2011). In all studies, except two, feature selection/reduction is applied. The number of subjects that participated in the studies also varies considerably, from 4 to 210. In contrast, the number of emotion classes that are to be discriminated is relatively similar among the studies (range: 5-8) with only one study that discriminates among 12 emotional states (Gunes & Piccardi, 2009). The reported recognition rates range from 72% to 98%. The early work of (Cottrell & Metcalfe, 1991) is an exception to this, with their 20% correct classification among 8 emotions (chance level: 12.5%). Taken together, classification performance varies strongly among different studies.

Audio-based emotion recognition. Audio-based emotion recognition mainly focuses on speech-based emotion recognition. In addition to speech-based emotion perception, studies have focused on non-linguistic utterances such as coughing, crying, laughing, and specific utterances such as ‘uuuuh’ (Petridis & Pantic, 2011). However, not all of these studies employed machine learning techniques. Moreover, the specific techniques employed for classifying these non-linguistic utterances results have not been adopted in speech-based

Table 3: Review of twelve machine learning studies employing speech to recognize emotions.

Reference	Ss	Feats	Functions	Targets	Perf
Lieberman & Michaels (1962)	3		10 humans	8 utterances	85%
Banse & Scherer (1996)	12		12 humans	14 emotions	48%
		18	LRM, LDA	14 emotions	25%
Nwe et al. (2003)	12		3 humans	6 emotions	66%
		64	HMM		78%
Oudeyer (2003)	6		8 humans	5 emotions	77%
				4 emotions	89%
		200	kNN, DT, ANN, K*, LRM, SVM, VFI, DT, NB	4 emotions	96%
Morrison et al. (2007)	11	38	SVM, kNN, ANN, NB, K*, DT	6 emotions	73%
Scherer et al. (2009)	10	200	kNN	7 emotions	70%
			20 humans		85%
Luengo et al. (2010)	10	383	SVM,GMM	7 emotions	78%
Tawari & Trivedi (2010)	10	1054	SVM	7 emotions	83%
	4	1054	SVM	3 emotions	88%
Sobol-Shikler & Robinson (2010)	10	173	DT,SVM	8 emotions	75%
Van den Broek et al. (2011)	25	65	LRM	SUD scale	83%
Wu et al. (2011)	10	442	SVM	7 emotions	92%

Notes. Abbreviations are explained in Table 1.

emotion recognition in general.

Similar to vision-based approaches, audio-based emotion recognition suffers from environmental noise (e.g., from a radio or conversations in the background). Moreover, audio recordings are influenced by acoustic characteristics of environments; however, using templates for distinct environments could relief this burden. Moreover, ubiquitous recording of sound could also cancel out some of the noise, although such an approach in itself is very challenging. Similar as with vision-based emotion recognition, audio-based emotion recognition is difficult when conducted outside a controlled setting as the measured signals contain more noise.

For decades audio-based emotion recognition has been conducted with a limited set of features (≤ 64), without the use of any feature selection or reduction; see Table 3. For representative reviews, we refer to K. R. Scherer (2003) and Ververidis and Kotropoulos (2006). An earlier review that should be mentioned is that of K. R. Scherer (1986), in which he discusses 39 earlier studies and 12 emotions and identified 18 features. During the last decade, the number of used features increased. Oudeyer (2003) used 200 features, which was at that time a huge number. The last three years, more often a brute force strategy was employed (Schuller, Batliner, Steidl, & Seppi, 2011), using hundreds or even thousands of features (Tawari & Trivedi, 2010; Wu, Falk, & Chan, 2011; see also Table 3). In parallel with the increase of the number of features used, feature selection/reduction strategies started to be applied in order to overcome overfitting and help selecting the most useful features.

Schuller et al. (2011) present a taxonomy of features for acoustic and linguistic emotion recognition, including: intonation (i.e., F0 or pitch), formants, intensity, Cepstral coefficients, spectrum, time-frequency transforms, harmonicity, perturbation, linguistics (e.g., phonemes and words), para-linguistics (e.g., laughter), and disfluencies (e.g., pauses). Ayadi, Kamel, and Karray (2011) provide a useful taxonomy as well, which shows significant resemblance with that of Schuller et al. (2011). These taxonomies provide a representative sample of features used in this research, such as those presented in Table 3. From each of these features a range of parameters can be determined, which make up the complete feature-parameter space. For an introduction on the speech emotion recognition processing pipeline, we refer to Ayadi et al. (2011).

It is noteworthy that in several studies professional actors participated who perhaps expressed emotions in a rather prototypical, not necessarily natural, manner. Moreover, it should also be noted that in speech-based emotion recognition, next to artificial classifiers, humans sometimes served as classifier. Speech-based human emotion recognition rates fall within the range of 50% – 75%, with 85% correct recognition reported by Lieberman and Michaels (1962) being an exception; see also Table 3. Machine recognition rates vary considerably with Banse and Scherer (1996), who report up to $40\%correct classification on 14 emotions, and Wu\ et al. (2011), who report 92\%$ correct classification on 7 emotions, and in between the results reported by Schuller et al. (2011) on the InterSpeech 2009 emotion challenge: $71\%(2 classes) and 44\%$ (5 classes).

Physiological emotion recognition. A broad range of physiological signals is used for emotion recognition; for example, cardiovascular activity (i.e., electrocardiogram, ECG and blood volume pulse, BVP), respiration, electrodermal activity (EDA), skin temperature, and facial muscle activity (i.e., fEMG), see also Table 4. The choice of the physiological signal(s) to be used depends on both the area of application (e.g., depending on which sensors can be unobtrusively incorporated in a device) and on the information that needs to be extracted from it. In addition, a delay between the actual emotional change and the resulting change in the recorded physiological signal has to be taken into account. Moreover, physiological signals differ significantly between individuals, which is why personalized approaches have been shown to provide the best performance (J. Kim & André, 2008). Furthermore, measurements with physiological sensors can be unreliable due to movement artefacts and differences in bodily position. More problematic, people's physiology is influenced by a variety of factors unrelated to emotions (Cacioppo & Tassinary, 1990), some with an internal (e.g., cognitive effort) and some with an external origin (e.g., a sound). Nevertheless, physiological recordings are gaining in popularity as the sensors become more and more sophisticated and their integration in other products can nowadays easily be realized. Physiological signal recording systems are now integrated in artefacts like helmets, beds, music players, gaming consoles, or clothes (Amft & Lukowicz, 2009; M. Chen, Gonzalez, Vasilakos, Cao, & Leung, 2011).

Table 4: Review of twelve machine learning studies employing different physiological signals to recognize emotions.

Reference	Signals	Ss	Feats	Function	Targets	Perf
Sinha & Parsons (1996)	M	27	18	LDA	2 emotions	86%
Picard et al. (2001)	C,E,R,M	1	40	LDA	8 emotions	81%
Scheirer et al. (2002)	C,E	36	5	HMM	frustration	67%
Kim et al. (2004)	C,E,S	175		SVM	3 emotions	73%
Lisetti & Nasoz (2004)	C,E,S	29		kNN, LDA, ANN	6 emotions	86%
Healey & Picard (2005)	C,E,R,M	9	22	LDA	3 stress levels	97%
Rani et al. (2006)	C,E,S,M,P	15	46	kNN, SVM, RT, BN	3 emotions	86%
Kim & Andr (2008)	C,E,M,R	3	110	LDA, EMDC	4 emotions	79%
Katsis et al. (2008)	C,E,M,R	10	15	SVM, ANFIS	4 affect states	79%
Chanel et al. (2009)	C,E,R	11	18		3/2 emotions	66%
	EEG		18720		3/2 emotions	73%
	C,E,R,EEG		18738		3 emotions	70%
Hosseini et al. (2010)	C,E,R	15	38	SVM	2 arousal	77%
	EEG	15	21	LDA, SVM	2 arousal	85%
Van den Broek et al. (2010)	E,M	21	10	kNN, SVM, ANN	4 emotions	61%

Notes. Abbreviations are explained in Table 1.

Few reviews have appeared on affective computing using physiological signals for emotion recognition, in particular when compared with the other modalities. This also illustrates that research on affective computing using this modality is rather new. It started with the publication of Picard's book Affective Computing in 1997 (Picard, 1997). At that moment, audio-based emotion recognition, in particular speech-based emotion recognition was already employed for decades. Recently, a concise review appeared (Knapp, Kim, & André, 2011), which briefly wraps up some key notions of physiological signals in affective computing. They report 92% correct classification rate as best result, using 4 signals and discriminating between 4 emotions.

Table 4 presents a review of key articles in this area of affective computing published throughout the last decade. In 2001, Picard et al. (2001) published their pioneering study with 81% correct classification on 8 emotions, also using linear discriminant analysis. Their study included multiple physiological signals but only one subject to which the complete classification processing pipeline was tailored. In the decade that followed this study, neither the classification rate of Picard et al. (2001), nor the number of emotions among which they discriminated were improved. However, the robustness of classification across participants has been addressed. J. Kim and André (2008) reported 70% correct classification on 4 emotions using a generic classifier and 95% correct classification when employing a personalized classifier. Most recent studies address generic classifiers that are applied using multiple participants; see Table 4. Most studies discriminated between 2-4 (categories of) emotions and achieved a correct classification in the range of 60% to over 90%.

Multimodal emotion recognition. As discussed, each of the three modalities predominantly applied in affective computing (i.e., vision, speech, and physiology) has its own specific advantages and disadvantages. However, in controlled environments and controlled experiments, each of these modalities separately has shown to provide a rich source of information on experienced emotions. As is illustrated in Table 5, so far the three modalities have not been combined for emotion recognition purposes. Vision and audio-based emotion recognition have been combined several times and even several review papers exist that take together both modalities, which illustrate their relation (Cowie et al., 2001; Lisetti & Nasoz, 2004; Pantic et al., 2011; Pantic & Rothkrantz, 2003; Zeng et al., 2009). Nonethe-

Table 5: Review of twelve machine learning studies employing multiple modalities to recognize emotions.

Reference	Signals	Ss	Feats	Functions	Targets	Perf
Tartter (1980)	Vi, Sp	6		12 humans	2 emotions	71%
De Silva et al. (1997)	Vi	2		18 humans	2 emotions	75%
	Sp					60%
	Vi, Sp				6 emotions	21%
Chen et al. (1998)	Vi	2	15	kNN, GMM	6 emotions	69%
	Sp		16			75%
	Vi, Sp		31			97%
Busso et al. (2004)	Vi, Sp	1	15	kNN, SVM	4 emotions	89%
Kim et al. (2005)	C, E, R, M	3	26	LDA	4 emotions	53%
	Sp		1280			52%
	C, E, R, M, Sp		1306			66%
Kim & Andr (2006)	C, E, R, M, S	3	77	LDA	4 emotions	51%
	Sp		61			54%
	C, E, R, M, S, Sp		138			55%
Kapoor et al. (2007)	Mo, Vi, Pre, E	24	14	kNN, SVM, GP	2 emotions	79%
Wang & Guan (2008)	Vi, Sp	8	153	GMM, kNN, ANN, LDA	6 emotions	82%
Chang et al. (2009)	C, E, T, Vi		15	ANN	4 emotions	95%
Petridis & Pantic (2011)	Vi, Sp	15	61	ANN	laughter	98%
Van den Broek (2011)	C, Sp	32	4	ANOVA	6 emotions	90%

Notes. Abbreviations are explained in Table 1.

less, physiological signals are hardly combined with other modalities. Only two groups have published on the combination of audio, in both cases it concerned speech, and physiological signals (J. Kim, André, Rehm, Vogt, & Wagner, 2005; Van den Broek, 2011; see Table 5). To our knowledge only two publications exist on the combination of vision and physiological signals (Bailenson et al., 2008; Kapoor, Burleson, & Picard, 2007).

The papers that do report multimodal affective computing incorporated few subjects. Exceptions are the works of Kapoor et al. (2007) and Van den Broek (2011) who included respectively 24 and 32 subjects, which is more in line with the studies on unimodal affect recognition (cf. Tables 2-4). The number of features extracted from the multiple signals differs significantly, ranging from 4 (Van den Broek, 2011) to more than 1000 (J. Kim et al., 2005). Surprisingly, the number of features seems to have little effect on the classification performance. Also, the number of emotions among which is discriminated is small, ranging from 2 to 6; see Table 5. This all suggests that work on multimodal affective computing is still in a beginning stage.

The classification results on multimodal affective computing are ranging between 21% and 98%. Most studies report recognition accuraries between 50% and 70% correct classification, which is slightly lower than the unimodal studies (cf. Table 5 with Tables 2-4). Most studies that do report higher classification rates have their drawbacks. For example, Petridis and Pantic (2011) report high classification rates, but they only discriminate between regular speech and laughter. An exception to this is L. S. Chen, Tao, Huang, Miyasato, and Nakatsu (1998) who report 97% correct classification among 6 emotion classes, using a combination of speech- and vision-based emotion recognition.

Conclusion. Despite the large body of work on machine emotion recognition (see Tables 2-5), the main conclusion of this review on affective computing has to be that it is very difficult to compare results of the different studies. This is because the studies differ in many aspects, including the number of participants and features used, the type and number of target emotions they try to recognize, and modalities they use. Therefore, studies with the highest recognition rates do not necessarily employ the best machine emotion recognition approaches.

Easier comparison between studies could be realized by creating a benchmark. A benchmark could help by identifying approaches to machine emotion recognition that work and approaches that do not work well. In turn, this would improve performance over time. Hence, the lack of standards and benchmarks (at least) partly explains the somewhat disappointing classification performances, when compared to other application areas of machine learning such as handwriting recognition, speech recognition, and image classification (in which there are standards and benchmarks). There, recognition performance shows a recognition rate of more than 90% and most often over 95%. Based on these results, it seems that we do not yet understand sufficiently how emotion recognition should be tackled.

Perhaps due to the complexity of unimodal emotion recognition, multimodal expression of emotions has been hardly touched upon. Few studies employed multimodal emotion recognition and some of the studies that did employ multimodal emotion recognition did not focus on a machine learning component (Bailenson et al., 2008; Van den Broek, Schut, Westerink, & Tuinenbreijer, 2009). For the case of unimodal emotion recognition, excellent review papers have been published that have drawn some conclusions on the progress of the field. In contrast, multimodal emotion recognition has not yet received similar attention. In order to keep the field of affective computing moving forward, we stress the importance of benchmarking the techniques employed and combining multiple modalities. Against this background, this paper will present a series of studies that create a fundamental benchmark, a direct comparison between man and machine using and combining the three modalities of speech, vision, and physiological signals.

Data gathering

To test how well humans and machines are able to recognize emotions, data is necessary to test them on. The first step of this research was a trial in which we collected self-reports of experienced emotions, physiological recordings, speech, and facial expressions from several people in various emotional states. As discussed before, it is key to generate ecologically valid emotional responses so that our results can be generalized to situations outside the lab. For this, we employed autobiographical recollection because (i) it is a very strong emotion inducer (Levenson, Carstensen, Friesen, & Ekman, 1991; Zaki, Weber, et al., 2009; Van den Broek, Van der Sluis, & Dijkstra, 2011), (ii) it is likely to generate more ecologically valid emotional responses compared to acted emotions that are sometimes used for emotion recognition research (e.g., Picard et al., 2001), and (iii) it involves speech whereas many passive induction techniques do not involve the participants having to speak (e.g., emotion induction through the presentation of emotionally charged pictures).

Experimental setup

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

Materials. Physiological recordings were done with a Mobi-8 of TMS International b.v. Electrodermal activity (EDA) was measured with two Velcro strips with dry electrodes strapped around the distal phalanx of the index and ring finger of the non-dominant hand. The EDA signal was sampled with 128 Hz. Skin temperature (ST) was measured with a thermistor strapped with medical adhesive tape to the distal phalanx of the little finger of the non-dominant hand. The ST signal was sampled with 128Hz. Respiration was measured with a gauge band positioned over the clothes around the chest at a sampling frequency of 128Hz. An electrocardiogram was taken at 1024Hz using Ag/AgCl electrodes on the standard Lead-II placement.

Facial expressions were recorded using a Logitec Quickcam webcam at 25fps with a resolution of 640 × 480 pixels. The webcam was positioned at the top of the computer screen that the participants were facing. Audio was recorded using a Sennheiser MZT 100 microphone positioned on the desk the participants were sitting at.

The entire experiment was done on a computer with the participants sitting behind a desk that held the computer screen, keyboard and mouse, microphone and webcam. The experimenter was in a control room next to the room the participants were in.

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

In the first phase of the experiment, participants were asked to recall two events from their personal experience that made them feel extremely happy, relaxed, sad, or angry. They were also asked to recall getting up this morning and going home last evening to provide two possibly neutral events. Subsequently, they were asked to write one paragraph about each of the ten events, shortly describing the event and their feelings during the event. Participants provided a title and rated each event on their own emotional intensity (very weak to very intense), valence (very unpleasant to very pleasant), and arousal (very relaxed to very aroused) on nine point Likert type scales during the event. This took 30 to 45 minutes.

In the second phase, the experimenter attached the physiological sensors and checked the signals visually. Subsequently, two individual difference measures were administered. First, the participant filled out the Berkeley Emotional Expressivity Questionnaire (Cronbach's $\alpha = .88$), a 16-item questionnaire assessing emotional expressivity (Gross & John, 1997). Second, we assessed participants' cardiac awareness (which is an objective index of emotional expressivity; Herbert, Pollatos, Flor, Enck, & Schandry, 2010) by having them count their number of heart beats without moving and with their hands side by side on the desk. Participants counted their heart beats over 25, 35, and 45 seconds without knowing the duration of these periods. Cardiac awareness was defined as

$$\frac{1}{3}\sum\left(1-\left(\left|recordedHBs-countedHBs\right|/\left(recordedHBs\right)\right)\right) \tag{1}$$

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

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

In the fourth and final phase, after all five events were described, participants were presented with (both audio and video of) each of the recordings we made of their disclosure. While watching the recording, participants had to continuously rate how they felt during the recording on a 9 point Likert type scale using the left and right arrow key. This is a validated method of emotional state assessment (Levenson, Ekman, Heider, & Friesen, 1992). Half of the participants first rated their arousal on all five movies and then their valence, whereas this was done vice versa for the other half of the participants. Valence and arousal are the two most commonly used dimensions for self-reported emotions (Lang, Greenwald, Bradley, & Hamm, 1993; Bradley & Lang, 1994).

Results

Duration. Duration of the disclosures of one of the participants was only 39 seconds, which was more than 3 SE from the mean duration of all participants. Therefore, data of this participant were discarded. The mean duration of the remaining disclosures was 141 seconds ($SE = 18s$). A repeated measures ANOVA (Huyn Feldt corrected to compensate for violation of the sphericity assumption) showed differences in duration between the five emotion types ($F(4, 32) = 4.664$, $p < .02$, $\eta^2 = .250$). Post hoc comparisons showed that relaxed ($M = 100s$) and neutral ($M = 85s$) emotion disclosures were shorter than angry ($M = 180s$) and sad ($M = 159s$) disclosures (p-values $< .05$). The duration of happy

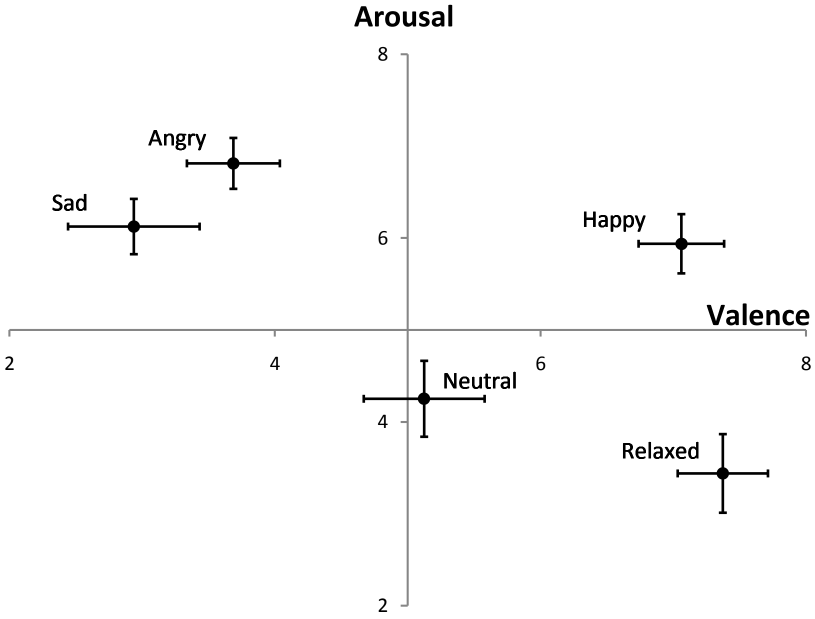

Figure 2. Mean arousal and valence ratings by the disclosers for each of the five emotion types. Error bars depict +/- 1 SE.

disclosures ($M = 102s$) did not significantly differ from the other durations (all p-values $> .05$). Furthermore, the duration of relaxed and neutral and the duration of angry and sad did not significantly differ (p-values $> .10$).

Self report ratings. To check if the emotion induction was successful we analyzed the self report ratings taken directly after each disclosure. Figure 2 presents the aggregated self-report ratings of emotional intensity, arousal, and valence, for each of the elicited emotions. A repeated measures ANOVA (Huyn Feldt) confirmed effects of Emotion type on Valence ($F(3, 46) = 24.847$, $p < .001$, $\eta^2 = .62$), Arousal ($F(4, 54) = 14.935$, $p < .001$, $\eta^2 = .50$), and Emotional Intensity ($F(4, 60) = 7.975$, $p < .001$, $\eta^2 = .35$). Pairwise comparisons showed that Emotional Intensity was lower in the neutral condition than in all the other conditions (all p-values $< .05$), whereas all the other conditions did not significantly differ in Emotional Intensity (all p-values $> .10$). Valence was higher in the Happy and Relaxed conditions than in all the other conditions (p-values $< .05$) and Valence was lower in the Sad and Angry conditions than in the other conditions (p-values $< .05$). Happy and Relaxed, and Sad and Angry respectively did not significantly differ on self-reported Valence (p-values $> .05$). Arousal was higher in the Angry condition compared to Happy, Relaxed, and Neutral (all p-values $< .05$) and there was a trend with Sad ($p < .07$). Happy and Sad where both higher in Arousal than Neutral and Relaxed (all p-values $< .05$) but did not differ from each other ($p > .10$). Neutral and Relaxed did not differ in Arousal ($p > .10$).

As an extra manipulation check, we also analyzed the means of the continuous ratings while participants reviewed the recordings of their own disclosures. These showed the same patterns as the overall ratings. A repeated measures ANOVA (Huyn Feldt) confirmed effects of Emotion type on Valence ($F(3, 45) = 30.383$, $p < .001$, $\eta^2 = .69$) and Arousal ($F(4, 56) = 8.972$, $p < .001$, $\eta^2 = .40$). Pairwise comparisons show that Valence was higher in the Happy ($M = 5.8$) and Relaxed ($M = 5.7$) compared to all other conditions (p-values $< .05$). Valence was higher in the Neutral ($M = 4.0$) condition than in the Angry and Sad conditions (p-values $< .05$). The Angry ($M = 2.9$) and Sad ($M = 2.2$) conditions did not differ in valence ($p > .10$). Arousal was lower in the Relaxed ($M = 2.6$) and Neutral ($M = 3.3$) conditions compared to all other conditions (p-values $< .05$). There was trend between Angry ($M = 5.1$) and Happy ($M = 4.3$) showing that Arousal in the Happy condition might be lower than Arousal in the Angry condition. Arousal in the Sad ($M = 4.6$) condition did not differ from the Angry and Happy conditions.

Individual differences. The results of the Berkeley Emotional Expressivity Questionnaire were calculated by averaging the positive items with the reverse coded negative items. The resulting expressivity scores had $M = 4.9$ and $SE = 0.24$ ($N = 17$). Because of the very limited range of expressivity scores it was not further used to analyze individual differences.

Cardiac awareness was calculated according to specifications given in the Procedure Section. With one exception score of .20, all the cardiac awareness scores were relatively high (all scores $> .60$) with $M = .76$ and $SE = 0.05$ ($N = 16$). Again, because of this limited range the cardiac awareness was not further used to analyze individual differences.

Conclusion

The employed emotion elicitation was successful. Participants disclosed around two minutes per emotion type, which we aimed for. Furthermore, the self-reported valence and arousal were congruent with the disclosed emotion, except for sadness. We expected sadness to be lower in arousal. The durations of the angry and sad disclosures were slightly larger than those of the other emotion types. Nonetheless, self-reported emotional intensity showed only a lower intensity for neutral. In sum, all four emotions were successfully elicited, with sadness being more arousing than was expected.

Human benchmark

To test how well humans are able to detect other's emotions, we ran two experiments in which humans watched the recordings of the disclosers and rated how they thought the disclosers felt. In both experiments, we compared all three combinations of the audio and video modalities (i.e., audio only, video only, and audio plus video) in the recordings to see how the communication modality influences emotion recognition performance. To allow for a fair comparison between human emotion recognition and machine emotion recognition we wanted to take out the semantic information in the recordings, as our computer algorithms also do not use semantic information contained in the speech. To that end, in the first experiment, we used participants that could not understand Dutch, but that did speak a language similar to Dutch so that the emotional prosody in the language was comparable. Therefore, we used American participants, as English is related to Dutch and therefore unlikely to contain strong differences in vocal affect expressions (K. R. Scherer & Zentner, 2001). Furthermore, we ran a second experiment in which we used Dutch participants, to also test how well they perform on the emotion recognition task with the semantic information available in addition to the non-verbal and para-linguistic features.

Non-Dutch Experiment: Experimental setup

Participants and Design. Participants were 22 female and 20 male undergraduates ($N = 42$) aged between 19 and 23 ($M = 20.4$) from an American west coast university. None of the participants knew any of the participants in the data gathering session and none of them spoke any Dutch. Participants received course credit for their participation.

Modality (audio / video / audio and video) was administered as a between-participant condition. Each participant received two blocks of five recordings. In one block they rated valence and in the other block arousal of the discloser. The order of the blocks was counterbalanced over the participants. Within each block there was a recording of each of the five emotion types (happy / relaxed / angry / sad / neutral). The order of the emotion types within each block was digram balanced (Wagenaar, 1969). For every participant, each of the ten recordings came from a different discloser.

Materials. Out of the 85 recordings made during the data gathering experiment, we selected a set of 40 recordings that had a variety of disclosers, relatively similar durations, contained each emotion type equally often, and had self-reported ratings of the discloser that were congruent with the emotion type. These constraints left 50 recordings from which we randomly selected a subset of 40. A subset was selected so that we would have the opportunity to collect multiple datapoints for one video while maintaining the emotional variety in the data that we gathered. In other words, each video would be presented to a number of different participants. We tried to select recordings with similar durations to limit the possible confound of different durations between the five emotion types.

Procedure. After consent for the experiment was obtained, participants saw pictures of all disclosers. Subsequently, depending on the condition, participants were instructed that they would see and/or hear ten recordings of persons talking about an event from their own lives in Dutch.

During the recording, it was the participants' task to pay careful attention to the recording. They also continuously rated either the valence or the arousal of the person in the recording using keyboard arrow keys. For the sake of clarity and brevity this data is not further reported. Directly after each movie we asked the participants to select the emotion that they thought was expressed in the recording. They could choose from five categories corresponding to the five categories used for the data gathering: happy, relaxed, sad, angry, and neutral. To make sure the participants did not understand the disclosers and did not use any semantic information for their judgments, we also asked them to describe what they thought the discloser was talking about. One participant sometimes understood what the discloser was talking about and was therefore removed from the sample. The entire experiment took about 30 minutes.

Results. To assess the emotion recognition performance we compared the emotions selected by the participants with the disclosed emotions. An ANOVA was run on the average percentage per participant run with Modality (Video only / Audio only / Video and audio) as between-participant factor. Results did not show a significant effect of Modality ($F(2, 39) = 1.18$; $p = .32$; partial $\eta^2 = 0.06$). Correct selection accuracy was 31.0% ($SE = 4.7\%$) in the Video and audio condition, 22.7% ($SE = 4.3\%$) in the Audio only condition, and 26.1% ($SE = 4.5\%$) in the video only condition.

Confusion matrices can be found in Tables 6, 7, and 8. From those confusion matrices, it can be seen that, in all conditions, Angry had the lowest correct predictions. In the audio-only condition, angry was mostly confused with happy and neutral. In the other two conditions, angry was mostly confused neutral and sad. In addition, happy was mostly confused with relaxed. Relaxed was mostly confused with happy and sad. Moreover, sad emotions were mostly confused with neutral, and neutral was mostly confused with angry.

We also compared the result of each condition against the 20% a priori probability of randomly guessing the correct emotion class using one sample one tailed t-tests. Recognition in the Video and audio condition is significantly higher than chance ($t(12) = 1.97$; $p < .05$). Recognition in the Video only condition is marginally higher than chance ($t(12) = 1.53$; $p = .075$). Recognition in the Audio only condition was not higher than chance ($p > .20$).

Dutch Experiment: Experimental setup

Participants and Design. Participants were 12 female and 15 male undergraduates ($N = 27$) aged between 19 and 25 ($M = 22.3$) from a Dutch university. None of the participants knew any of the participants in the data gathering session and all of them were native Dutch speakers. Participants received 5 Euro for participation. The materials and procedure were the same as in the Non-Dutch experiment. Participants were equally divided over the three modality conditions: video only, audio only, video and audio.

Table 6: Confusion matrices for human emotion recognition of the American samples for Audio only. Average correct recognition rate is 22.7%.

	Predicted emotion				
Disclosed emotion	Happy	Relaxed	Sad	Angry	Neutral
Happy	25.9%	47.4%	10.4%	15.6%	0.0%
Relaxed	14.1%	25.1%	28.6%	17.6%	14.1%
Sad	24.4%	12.0%	24.4%	9.0%	30.3%
Angry	50.0%	0.0%	0.0%	0.0%	50.0%
Neutral	16.4%	12.8%	20.0%	29.0%	21.8%

Table 7: Confusion matrices for human emotion recognition of the American samples for Video only. Average correct recognition rate is 26.1%.

	Predicted emotion				
Disclosed emotion	Happy	Relaxed	Sad	Angry	Neutral
Happy	32.0%	28.2%	0.0%	32.0%	7.7%
Relaxed	28.0%	25.0%	18.5%	12.5%	15.5%
Sad	13.2%	13.2%	34.7%	17.4%	21.6%
Angry	0.0%	16.1%	25.3%	16.1%	41.4%
Neutral	17.4%	17.4%	21.6%	21.6%	21.6%

Table 8: Confusion matrices for human emotion recognition of the American samples for Audio and video. Average correct recognition rate is 31.0%.

	Predicted emotion				
Disclosed emotion	Happy	Relaxed	Sad	Angry	Neutral
Happy	28.0%	34.4%	9.2%	12.4%	15.6%
Relaxed	28.2%	32.3%	24.1%	8.2%	8.2%
Sad	19.2%	7.9%	31.0%	15.3%	27.1%
Angry	20.5%	10.3%	39.7%	20.5%	10.3%
Neutral	5.9%	11.4%	14.3%	37.4%	31.5%

Results. To assess the emotion recognition performance we compared the emotions selected by the participants with the disclosed emotions. An ANOVA was run on the average correct percentage per participant with Modality as between-participant factor. Results showed a significant main effect of Modality ($F(2, 25) = 8.29$; $p < .002$; partial $\eta^2 = 0.43$). Correct selection was highest in the Audio only condition, with an accuracy of 62.8% ($SE = 4.9\%$). The correct selection in the Video and audio condition was 47.9% ($SE = 4.4\%$). Correct selection was lowest in the Video only condition, with an accuracy of 35.7% ($SE = 5.2\%$). Pairwise comparisons show a significant effect when comparing the audio only with the video only condition ($p < .005$). Marginally significant effects for the comparison of the Video and audio condition with the Video only condition ($p = .09$) and with the Audio only condition ($p = .06$).

Confusion matrices can be found in Table 9, 10, 11. Across all three conditions Neutral and angry had the lowest prediction rates. Neutral was least often confused with happy, but was confused with all other emotion classes. Angry was mostly confused with sadness. Happy was mostly confused with relaxed, and relaxed was mostly confused with neutral.

We also compared the result of each condition against the 20% a priori probability

Table 9: Confusion matrices for human emotion recognition of the Dutch participants for Audio and video. Average correct recognition rate is 47.9%.

	Predicted emotion				
Disclosed emotion	Happy	Relaxed	Sad	Angry	Neutral
Happy	46.7%	33.3%	13.3%	0.0%	6.7%
Relaxed	14.3%	64.3%	0.0%	7.1%	14.3%
Sad	10.0%	0.0%	90.0%	0.0%	0.0%
Angry	16.7%	0.0%	33.3%	25.0%	25.0%
Neutral	7.7%	38.5%	23.1%	7.7%	23.1%

of randomly guessing the correct emotion class using one sample one tailed t-tests. Video and audio condition ($t(7) = 12.78$; $p < .001$), Video only condition ($t(9) = 2.60$; $p < .02$), and Audio only condition ($t(7) = 11.61$; $p < .001$) all result in significant differences with 20% a priori probability. Finally, we also compared the Video only condition of the Dutch sample with the Video only condition of the Non-Dutch sample, as these were expected to lead to similar results. As expected, no significant difference was found on a one sample t-test with a reference value of 26.1% ($t(9) = 1.61$; $p > .10$).

Discussion

From the analyses of the Non-Dutch sample it becomes clear that Audio-and-video results in the best human emotion recognition performance, followed by the video-only conditions, and with the worst performance for the audio-only conditions. This makes sense, considering that audio and video together contains the most information and is also a more common way of emotion communication than communication through either visual or audio modalities only. Nonetheless, it is also clear from these findings that the recognition performances are only slightly higher than random guessing. The Audio-only condition does not even show significant differences with random guessing. Hence, this suggests that emotion recognition is a difficult task when no semantic or context information is present.

Table 10: Confusion matrices for human emotion recognition of the Dutch participants for Video only. Average correct recognition rate is 35.7%.

	Predicted emotion				
Disclosed emotion	Happy	Relaxed	Sad	Angry	Neutral
Happy	39.1%	21.7%	8.7%	4.3%	26.1%
Relaxed	8.7%	34.8%	21.7%	0.0%	34.8%
Sad	0.0%	6.3%	50.0%	12.5%	31.3%
Angry	10.0%	10.0%	40.0%	30.0%	10.0%
Neutral	21.1%	10.5%	15.8%	26.3%	26.3%

Table 11: Confusion matrices for human emotion recognition of the Dutch participants for Audio only. Average correct recognition rate is 62.8%.

	Predicted emotion				
Disclosed emotion	Happy	Relaxed	Sad	Angry	Neutral
Happy	66.7%	14.3%	4.8%	4.8%	9.5%
Relaxed	15.0%	55.0%	5.0%	0.0%	25.0%
Sad	0.0%	0.0%	78.6%	7.1%	14.3%
Angry	5.6%	0.0%	27.8%	55.6%	11.1%
Neutral	5.9%	5.9%	23.5%	29.4%	35.3%

In the Dutch sample a different pattern emerges: as expected, the audio is much more valuable because for this sample it also contains semantic information about the emotions described by the disclosers. Hence, it makes sense that the Audio-only condition leads to a much higher performance than the Video-only condition. Interestingly, the audio and video condition results in a lower performance than the audio-only condition. This suggests that more information does not necessarily lead to a better performance. Instead, it might be the case that participants were less focused on the speech or were thrown off by a perceived mismatch between the video and audio channels.

In previous work that compared audio and video recordings of emotions, participants spoke the same language as used in the recordings (Zaki, Bolger, & Ochsner, 2009). Therefore, audio conditions contained not only emotional prosody information but also semantic information because of what was told in the recording. We removed the effects of semantic information in the speech by using participants that did not speak the language used in the recordings. This way, the participants got the emotional information from the speech but not the semantic information of the stories that were told in the recordings. This allowed for a more direct comparison between human and machine recognition. Furthermore, it investigated emotional speech independent of the semantic information, which might, in practice, be more realistic as people can have emotions while not talking about those emotions but about something else (e.g., in case they want to hide their emotion). Taking the results of the Dutch and American samples together, it is seen that the value of the audio channel is mainly because of its semantic information. However, when audio is purely used for the emotional prosody, facial expressions are more informative emotion recognition channels for humans than speech.

There are some limitations to our approach. First of all, differences between the two samples might also be due to cultural effects. This is unlikely for the effects of the facial expressions as these tend to be universally the same (Ekman & Friesen, 1971). Moreover, there have also been researchers that have suggested that affective speech in related language like Dutch and English is very similar (K. R. Scherer & Zentner, 2001). In addition, we are unaware of a method for separating the emotional prosody of speech from the semantic information. Moreover, this is also one of the reasons why we added the Dutch sample as an extra comparison group with machine emotion recognition.

Second, the emotional expressiveness of the participants in the data gathering session might have been limited because they were not in a regular social context. When interacting directly with others, they might show more of their emotions through facial expressions or speech intonations (Fridlund, 1991; Fischer, Manstead, & Zaalberg, 2004). Moreover, because we used movies, the participant who had to recognize the emotions could not influence or interact with the recorded persons. This might have limited their capability of recognizing the emotions. We had, however, two reasons to take this approach. First, we aimed to create a fair comparison with machine emotion recognition. Machine emotion recognition is often done in a non-social context with no interaction between the machine and the human. Moreover, we used a relatively controlled experimental setup to be able to induce similar emotions in all participants and to be able to show the same stimuli to participants in different recognition conditions. This is an approach taken by other emotion recognition studies as well (Ickes & Aronson, 2003; Ickes, 1997; Zaki, Weber, et al., 2009). Finally, video has limitations (e.g., in terms of resolution and depth perception) compared to face-to-face interaction, that might have decreased emotion recognition performance. However, we used video for the in human emotion recognition experiments as we aimed to create a fair comparison between machine and human emotion recognition.

Given the limitations discussed above, it is clear that human emotion recognition performance is not very high in absence of semantic or context information. In such cases, it seems that people make a correct prediction of how someone else is feeling in only one third of the cases if they have to select from five emotions. This shows that the task of emotion recognition solely from facial and vocal expressions is difficult. In situations with semantic information, humans perform much better. Hence, it might be worthwhile for future studies to include semantic information (and context) as an important factor for both human and machine emotion recognition. Nonetheless, the resulting recognition accuracy of 62.8% for a five class problem is still relatively low when comparing it to other tasks in which humans perform much better. In sum, these findings suggest that emotion recognition is a difficult problem compared to other pattern recognition domains.

Machine emotion recognition

In this section, we describe the results of training a machine on the data that we gathered and see how well the trained machine performs on the emotion recognition task. We extracted features from the video, audio, and physiological modalities as input for the training from the same dataset as was used to test human emotion recognition. Subsequently, we let the machine predict the class of emotions from which the features were extracted using two different statistical machine learning techniques. We first describe the method in detail and will then present the classification results.

Method

The machine emotion recognition was investigated using video, audio, and physiological data from the participants in the data gathering experiment to investigate whether the machine could classify this data into the five classes of emotions induced in the data gathering session. First, features were extracted for each modality separately. Second, for the physiology features baseline correction was applied, as their baselines can change over time. Third, modalities were fused using feature-level fusion (i.e., by concatenating feature vectors of the different modalities). Fourth, feature selection was applied in order to find the most useful features for separating the emotion classes. Finally, classification was done for each modality separately and for every combination of the three different modalities, using a support vector machine (SVM) and a multilayer perceptron (MLP).

Feature extraction. For feature extraction, a running window of 10 seconds was used, over which features were extracted. Each step, the window was shifted 10 seconds, preventing subsequent time windows from overlapping. A 10-second window was chosen as this relates to the speed with which emotional expressions can change (Cowie et al., 2001).

Facial expressions were analyzed using the OKAO Vision software package from OMRON Corporation (Ahn, Bailenson, Fox, & Jabon, 2010). This software tracks 37 points on the face, primarily around the mouth and eyes and general features of the head like mouth and eye openness, and the yaw, pitch, and roll of the head. From the extracted features, motion features were generated by computing the difference between two sequential feature values (Fasel & Luettin, 2003). These values were extracted at 15 frames per second and then averaged for every ten seconds of data. This gave 42 facial feature values for each ten seconds of data.

Audio was obtained from the recorded videos. For extracting audio features, the software PRAAT (Boersma & Weenink, 2010) was used. The features we extracted were based on features that have often been used in the literature. First of all, we extracted pitch, which is often referred to as fundamental frequency F0 and represents vocal-cord vibration as a function of time (Boersma & Weenink, 2010). Pitch can, for example, be used to distinguish happiness or sadness from a neutral state (Cowie & Cornelius, 2003). We also extracted intensity, which represents the speech signal intensity in dB (Boersma & Weenink, 2010). Third, formants describe vocal resonances at specific frequencies in the frequency spectrum. Each formant has its own frequency range. The use of formants has also been found useful for emotion classification (Ververidis & Kotropoulos, 2006). Nygaard and Lunders (2002) linked sadness to increased jitter and shimmer values, which we extracted as well: jitter is a term for the frequency variability of the pitch signal and shimmer describes the amplitude variability of the pitch signal (Boersma & Weenink, 2010). The fraction of unvoiced frames in the pitch signal was also used as a feature value (Busso et al., 2004), in which each frame has a duration of 0.01 seconds. Finally, we used the degree of voice breaks, which is the relative duration of breaks between the voiced frames (Boersma & Weenink, 2010). This resulted in 12 audio features, namely: mean and standard deviation of pitch, mean and standard deviation of intensity, formants F1 to F4, jitter, shimmer, the fraction unvoiced frames, and the degree of voice breaks. These are features that are commonly used in machine emotion recognition studies (Schuller et al., 2011).

The physiological signals from which features were extracted were electrodermal activity (EDA), skin temperature (ST), respiration (RSP), and electrocardiogram (ECG). EDA and ST were preprocessed with a 0.5Hz low pass filter and downsampled to 2Hz. Subsequently, mean, standard deviation, and slope were calculated over each 10 second window. For the EDA signal, we also extracted the number of skin conductance responses using the SCRGauge algorithm (Boucsein, 1992). The respiration signal was filtered with a low pass filter (cut-off frequency: 0.5Hz) before individual breaths were extracted using a local min/max filter based on the method of Lemire (2006), from which the respiration rate was derived. From the ECG signal, first the inter-beat intervals (IBIs) were detected using pattern matching. Subsequently, mean, standard deviation, and root mean squared successive differences were calculated as time domain heart rate variability measures (Berntson et al., 1997). The IBIs were also transformed to the frequency domain from which the power in the low frequency (0.05Hz - 0.15Hz), high frequency range (0.15Hz - 0.40Hz), and the ratio between low frequency and high frequency power were calculated (Berntson et al., 1997). For these calculations we used 30-second windows instead of 10-second windows as the lower frequency components have cycle times over 10 seconds. Thirty seconds is thought to be the minimum duration to get reliable heart rate variability (HRV) estimates (Porges et al., 1997). In total we derived 14 physiological features.

This resulted in 715 feature vectors. Of all the feature vectors, 135 (19%) had a happy class, 110 (15%) had a relaxed class, 179 (25%) had a sad class, 195 (27%) had a angry class, and 96 (13%) had a neutral class.

Preprocessing and normalization. Baseline correction was applied to physiological signals as physiological baseline values may change over time. During the data gathering, each emotion recording was preceded by a 5 minute baseline. We extracted the mean and standard deviation from the last 90 seconds of this baseline period, to make sure effects of the previous emotion disclosure were not included in the baseline values. The features from the subsequent emotion recording were baseline corrected using a standard z transformation, by subtracting the baseline mean of the entire trace from each datapoint and dividing each point by the baseline standard deviation (Boucsein, 1992).

Finally, all the features from audio, video, and physiological data were scaled in $[-1, 1]$, by subtracting the minimum value and dividing by the difference between the maximum and minimum value of each feature. This was done to prevent features with large numeric values to dominate with smaller numeric values (C.-W. Hsu, Chang, & Lin, 2003).

Feature selection. Not all features contribute equally to dividing the different emotion classes. In order to find the features that best separate the five emotion classes, feature selection was applied. Two feature selection methods were used. The first method was to select (combinations of) the different modalities by hand. In this case, all features from a modality were either included or excluded in the classification process. This way, we compared all different modalities and all combinations of modalities with each other, leading to seven different combinations.

The second feature selection method was an automated method using a wrapper approach for finding the best feature subset out of the features of all modalities combined. A wrapper is constructed of a search algorithm for finding the optimal subset and a classification algorithm for evaluating the quality of each subset (Kohavi & John, 1997). The type of classification algorithms were the same as used in the final evaluation (see next section). Moreover, as in the final evaluation, the classification algorithms employed fivefold cross validation (see next section). The search algorithms that were selected were best first search and genetic search (S. Russell & Norvig, 1995). In short, best first search adapts the feature set by adding or removing a feature of the feature set that performed best up to that moment and evaluates the new set. Genetic search makes random adjustments (mutations) to the feature set, evaluates those adjustments against the non-adjusted set to, and continuous with the best performing set of features.

Classification. Two classifiers were used: a support vector machine (SVM), with the radial basis function (RBF) kernel, and a multilayer perceptron (MLP). These two techniques were selected based on their successful application in a wide variety of applications (Bishop, 1995, 2006). Furthermore, as shown in Tables 2 to 5, they constitute two of the most popular techniques used for classification of emotions. For the SVM, the LIBSVM implementation (C.-C. Chang & Lin, 2011) was used, with parameters cost and gamma values set to 100 and 0.1 respectively. For the MLP, the WEKA implementation with its default settings was used (Witten & Frank, 2005). The number of hidden layers is 1. The number of neurons in the hidden layer was determined by rounding down the result of $(N_f + N_c)/2$ where N_f depicts the number of features and N_c depicts the number of classes. In addition, a sigmoid activation function was used. Learning rate was set to 0.3, momentum was set to 0.2, and 500 epochs were used for training.

Cross-validation was done using five folds. Hence, the data was randomly split into five equally sized parts. Five runs were done and in each run a different part was selected for testing while the classifier was trained on the data in the other four parts. Finally, the

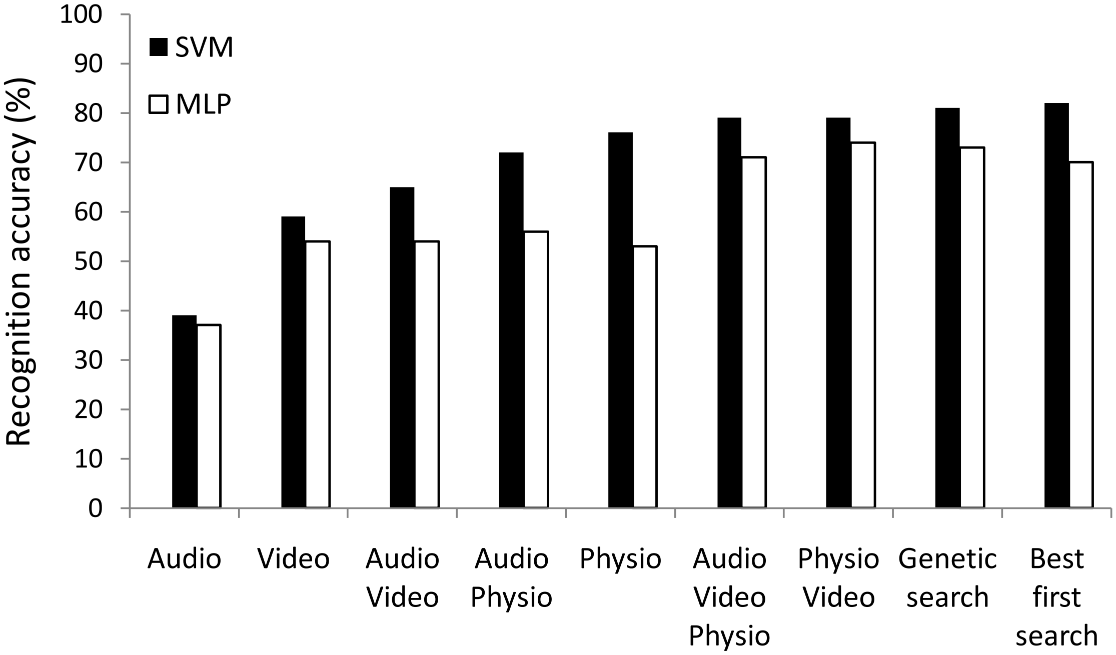

Figure 3. Recognition accuracy for the different methods of machine learning techniques. The x-axis depicts the different combinations of features and modalities. The y-axis depicts the recognition accuracy as obtained by using fivefold cross-validation. Bars for the support vector machine results are black and bars for the multi layer perceptron results are white.

results of testing recognition accuracies were averaged.

Results and Discussion

We trained the classifiers over the data of all participants, using SVM and MLP classifiers with fivefold cross-validation. Different feature subsets were selected in two ways: 1) based on different sets of the modalities (i.e., video, audio, and physiology) that we used and 2) based on wrapper selection techniques using genetic search, and best first search. The overall results of the recognition accuracies of all these different methods are presented in Figure 3.

Comparing the different modalities shows that audio alone performs worst with 39% (SVM) and 37% (MLP) recognition accuracies. Video performs better with 59% (SVM) and 54% (MLP) recognition accuracies. Combining audio and video features results in 65% (SVM) and 54% (MLP) recognition accuracies. Based on the SVM, the physiologi-

Table 12: Classification results by leaving out different physiological sub-modalities using support vector machines. Leaving out RSP performs the best whereas leaving out ECG performs the worst.

SC	ST	RSP	ECG	Recognition rate
x	x	x		59.58%
x	x		x	77.62%
x		x	x	65.73%
	x	x	x	67.41%
x	x	x	x	75.94%

*S*C: Skin Conductance; ST: Skin temperature; RSP: Respiration; ECG: Electrocardiogram.

cal features outperform the audio and video features with 76% (SVM) and 53% (MLP). Combining physiological and audio features results in 72% (SVM) and 56% (MLP). For the SVM, this is lower than using the physiological alone, suggesting that the audio feature actually introduce noise making it more difficult to train the classifier. Combining physiological features with video features shows an increase in recognition performance to 79% (SVM) and 74% (MLP). This indicates that physiological information has potential for improving machine emotion recognition of emotion. Moreover, although adding speech does not increase the performance. Instead, it is worthwhile to think about systems that can combine both physiological and facial expression information as this yields the highest emotion recognition performance.

We further investigated the contribution of different physiological modalities by classifying emotions using SVMs and selecting specific sets of features of the four physiological modalities. The results of this are depicted in Table 12. Classification with all features gives a classification performance of 76% whereas classification with the respiration features left out gives a performance of 78%. This suggests that respiration features are the least useful, as including them actually decreases classification performance. This might be because they are confounded by the speaking of the participants, which influences the respiration rate. Hence, the respiration features mainly add more noise to the data. The ECG features contribute the most to the performance. Leaving them out decreases the recognition accuracy to 60%. Leaving out the skin temperature or the skin conductance features also decreases the recognition accuracy, but not as much as leaving out the ECG features.

We also analyzed the results of the second method for feature selection: a wrapper method with genetic search and best first search. This turned out to lead to a slightly higher recognition accuracy compared to previous analysis (see Figure 3). The best first search in combination with an SVM led to the highest classification performance of 82%. The final set of selected features is presented in Table 13. A confusion matrix of the resulting classification can be found in Table 14. In line with the other findings, no audio features were selected. Moreover, also no respiration features were selected. The physiological set consists mostly of cardiac features. This is in line with the psychological literature that has identified heart rate and heart rate variability as an important measure of emotions (Berntson, Cacioppo, & Quickley, 1993; Butler, Wilhelm, & Gross, 2006; Grossman & Taylor, 2007; Kreibig, 2010). The facial feature set consists mainly of features around the eye and mouth. In sum, the results of both feature selection techniques we applied are clearly aligned. They both show that audio is the least useful, and that the best performance is obtained when combining physiological and facial signals. Finally, the SVM outperforms the MLP on all feature sets. To our knowledge, we are the first to compare all three modalities, so it remains to be seen how these results generalize to other datasets.

These results illustrate the advantages of combining different modalities for machine emotion recognition. As discussed in the background section, there have not been many studies that have studied the combination of different modalities. As expected, machine emotion recognition was better when combinations of modalities were used. As we used a controlled laboratory environment, the benefits of modality fusing might be even more pronounced in real-world settings. In such contexts, measurements are less reliable and

Table 13: The features selected for the best classification performance. The features were selected using best first search. Audio features were included in the search space but not selected in the best performing set.

Physiological features	Video features
Mean SC	Confidence
Number of SC responses	Standard deviation of face
Mean ST	Right eye ratio
Standard deviation of ST	Mouth ratio
Slope ST	Yaw
Mean HR	Right pupil y-value
LF power HRV	Left pupil y-value
HF power HRV	Lower lip center y-value
RMSSD HRV	Right mouth corner x-value
SDNN HRV	Right mouth corner y-value
	Left mouth corner x-value
	Left mouth corner y-value
	Left outer eye corner y-value
	Right outer eye corner x-value
	Left lower lip y-value
	Gaze tilt
	Gaze pan
	Left eye openness
	Mouth openness

*S*C: Skin Conductance; ST: Skin temperature; HR: Heart rate; HRV: Heart rate variability; LF: Low frequency (0.05-0.15Hz); HF: High frequency (0.15Hz-0.40Hz); RMSSD: Root mean squared successive differences; SDNN: Standard deviation IBI intervals.

Table 14: Confusion matrix for the classification results of the support vector machine with the best first search, which was the best classification method with 82% recognition accuracy. Rows depict the emotion predicted by the computer and columns depict the actual disclosed emotion.

	Disclosed emotion				
Predicted emotion	Happy	Relaxed	Sad	Angry	Neutral
Happy	78.52%	7.27%	1.68%	4.1%	9.38%
Relaxed	6.67%	79.09%	1.68%	2.05%	2.08%
Sad	6.67%	5.45%	88.83%	4.1%	8.33%
Angry	5.93%	7.27%	4.1%	86.67%	10.42%
Neutral	2.22%	0.91%	8.33%	3.08%	69.79%

audio or video signals could be degraded due to sensor or bandwidth limitations making multimodal verification more important (Van den Broek, Janssen, Westerink, & Healey, 2009). Therefore, we think it is worthwhile to further investigate multimodal emotion recognition by, for instance, comparing different fusion and synchronization techniques.

Another point that could further improve the machine emotion recognition is to include temporal dynamics in the model (Van den Broek, 2010). In our analyses, we have dealt with this by taking short time windows that are related to the temporal resolution of emotions (Cowie et al., 2001). However, subsequent time windows are unlikely to be completely independent from each other. For instance, the law of initial values (Wilder, 1967) suggests that physiological changes are based on the current physiological level (e.g., heart rate is more likely to decrease than increase when it is very high). Moreover, emotions influence longer term moods. In turn, moods influence emotions. This way, specific emotions can also have longer term affective influences, and subsequent emotions might be related. Possible models to account for such temporal relationships are dynamic Bayesian networks (Bishop, 2006) or time series analyses like autoregressive or moving average models (Box & Jenkins, 1976).

General discussion

It was our goal to compare human and machine emotion recognition on the same dataset using facial expressions, speech, and physiological signals. For this, using autobiographical recollection, we elicited five emotional states in participants while we recorded their speech, facial expressions, and physiological signals. These recordings were shown to others who had to estimate which of five emotions had been induced in the participant in the recording. Subsequently, we used the same data to conduct machine emotion recognition experiments with different statistical learning methods. The detailed results have been discussed in the previous sections, so we will limit ourselves here to the key findings of our studies.

Machines outperform humans

When comparing the results of human and machine emotion recognition studies, we can see that the machines performed better than humans. Although we have seen many studies attempting to train machines to recognize emotions, the recognition performances of these studies have been difficult to interpret. We show that machines perform much better than humans in situations where they are given access to the same signals. Across all modalities and the two samples we used to test human performance, the machines outperformed the human recognition accuracy.

There are several potential reasons for the fact that the machines outperform human emotion recognition. First, the machines have different information at their disposition. Machines have access to physiological signals that humans are not able to perceive (without the use of devices). Moreover, the features extracted from the modalities for the machines might be more useful for emotion recognition than the features that humans use. This could be the case because machines have more time to detect the features, whereas humans might have evolved to optimize the speed with which they can detect emotions, as emotions are often used as a warning system (Damasio, 1994). Furthermore, the machines might outperform humans because humans are geared to detecting emotions expressed in a social context. However, many affective computing applications are done outside a social context, which led us to use data that was not gathered in a social context. The machines might be better at handling such data, as they are trained specifically to do so. Finally, a lack of motivation for accurately recognizing the emotions might have led the humans to perform less accurate (Ickes, Gesn, & Graham, 2000). In contrast, for machines this is not an important issue. In sum, machines have several benefits over humans for recognizing emotions that can explain their performance differences.

A few other studies have also made comparisons between human and machine classification. For instance, Lien et al. (2000) showed that vision based classification is comparable to the inter rater agreement of two human facial action coding specialists. However, they focused on comparisons between computer and human experts and not on emotion recognition by lay people as we did. In addition, Nwe, Foo, and Silva (2003) report that machine classification algorithms outperform human classifiers using speech based emotion recognition. However, they relied on an acted emotion corpus and used only three humans for recognition. Although those studies were different from ours and focused on comparisons of only one modality, their findings are in line with our results. All in all, this shows that machine emotion recognition outperforms human emotion recognition for various modalities.

Other studies have focused specifically on recognizing human facial expressions. The studies of Ekman and Friesen (1971) and Ekman and Friesen (1986), in which it is shown that specific facial expressions can be recognized universally with very high accuracy, can be considered as an example of this. This research was, however, based on very clear static photos, instead of the more ecologically valid ambiguous dynamic disclosure recordings that we used. Other studies that have used recordings similar to ours have also shown that humans are far from perfect at recognizing other's emotional states (Ickes, 1997; Zaki, Weber, et al., 2009; Zaki, Bolger, & Ochsner, 2008). Hence, this suggests that emotion recognition is a challenging task.

It is good to note that it is very difficult to create a precise one-to-one comparison between the human and machine emotion recognition conditions. Points in which the human and machine recognition differed are outlined below. First, the human recognizers did not know the participants in the dataset and did not get the chance to get to know them. In contrast, the machines were trained to learn to recognize the emotions from these specific persons. Second, the humans in our experiment have emotion recognition skills that generalize to a larger population than those of the machines that are specifically tailored to the set of persons used in our experiment. Third, the humans have had much more time and data to train their skills compared to the machines. Therefore, in the future, it is worthwhile to see how our methods and findings hold for other corpora. All in all, we tried to create ecologically valid conditions to have a fair comparison between human and machine emotion recognition.

To create a fair comparison, part of the people that expressed the emotions spoke a different language from the ones that did the emotion recognition. We did this on purpose, as machines cannot easily benefit from understanding the semantics and context of spoken language. Moreover, we want to be able to generalize our emotion recognition to the many situations in which semantic information from speech is absent. Therefore, as an extra experimental control, we also included a second sample that did speech Dutch. Those participants did perform better when they had the semantic information of the speech, but did still worse than the machine emotion recognition. Second, we used Dutch disclosers and English recognizers as Dutch and English are related languages and emotional prosody is similar in those languages (K. R. Scherer, Banse, & Wallbott, 2001). Hence, a cultural difference in the interpretation of prosody would have been unlikely. This is also confirmed by similar recognition performances for the video-only condition between the two samples. Finally, the three conditions (audio, video, and audio plus video) show the same recognition pattern for the non-Dutch as well as for the machine recognition. This suggests that the two modalities had similar relative value for the humans as for the machines.

We have explored person-independent classification of emotions instead of person-specific classification. This was done because it relates to the wide range of applications in which emotion recognition has to be used without calibration to a specific user. Moreover, it was in line with the human emotion recognition as the human recognizers also did not know the disclosers. Nonetheless, we know from psychological studies that humans become better at recognizing other's emotions when they become more familiar with those (Ickes, 1997). Moreover, machine emotion recognition studies have confirmed these findings by showing that person-specific models improve recognition accuracy (J. Kim & André, 2008; Bailenson et al., 2008). Hence, it might be worthwhile to further investigate the combinations of different modalities, and compare human and machine recognition in settings where both are allowed to become familiar with the discloser.

The fact that machine emotion recognition outperformed human emotion recognition leads to a few interesting points for future research. First of all, it suggests that the models for machine emotion recognition we currently use are promising for real-world applications. Although most researchers motivate their research with affective computing applications, only few researchers actually develop and validate such applications. Now that we know that machine emotion recognition is at a high level compared to human performance, we believe that it is important to start transferring the successful emotion recognition models into applications and see how well they perform in practice. Moreover, this also includes testing how well the machine learning algorithms perform out of the lab. Currently, only very few studies are conducted in real-world contexts. However, the real-world contexts contain more noise than laboratory environments and actual real-world performance is therefore likely to be lower in practice for the machine emotion recognition. In contrast, humans might perform better in real-world contexts because in those cases they can directly interact with the persons they are trying to recognize the emotions from. In sum, real-world experiments should be conducted to further validate the machine's performance for actual applications. The performance of the machine emotion recognition can then be based on the performance of the application. For instance, a music player aimed to direct mood could be evaluated against its success in directing mood. A car signaling stressful moments to the driver could be evaluated based on the number of accidents that it can avoid.

Humans get a lot of information from context and semantics of the speech. Hence, to further improve machine emotion recognition performance, it might be worthwhile to start including semantics and context in emotion recognition models. This could, for instance, be done by employing speech recognition to analyze the semantic context of the speech. The potential usefulness of considering the context becomes apparent when one views emotion detection as a form of abductive reasoning (Peirce, 1933). Basically, an observer of another person generates a hypothesis to explain all incoming information in the most coherent way possible. Abduction in realistic circumstances is a notoriously difficult computational process (Fodor, 1983; Pylyshyn, 1987; Haselager, 1997). In principle many different hypotheses can be generated to explain the behavior (including physiological measurements) of a person, e.g. a specific raising of an eyebrow could be taken as a simple muscle twitch, a sign of exasperation, surprise, or increased attention, etc. Normally, the context can provide cues to direct the generation of hypotheses (as well as the selection of the best) about what particular emotion a person is experiencing. Particularly, aspects of the (wider) current environment, previous interactions, the purpose of the particular meeting, and even the goals of the observer may help to limit the set of potentially relevant hypotheses. Therefore, a useful approach might be to extract emotion from stimuli that a person encounters (e.g., movies) to better predict how someone reacts to that stimulus. Similarly, using information about the main purpose of the interaction with the other person could substantially facilitate emotion recognition. The reason why you are interacting may help to focus on a subset of emotions and their intensities, give clues as to what people might try to show or hide, and to know what kind of cues would be (in)appropriate for the situation. Moreover, one's own interests provide a framework that influences the kind of explanatory hypotheses one is (un)likely to generate. Incorporating these approaches could further improve machine emotion recognition.

Physiology is a promising modality

Our second key finding is related to the comparison of different modalities. Our results show that physiological signals are the most useful for recognizing emotions. Within the physiological modalities, electrocardiogram is the most valuable and respiration is the least valuable. Facial expression analysis improves recognition performance when added to the physiological data. This is also confirmed by automated feature selection, which gives the same results as comparing the modalities one by one. Now that all kinds of wireless unobtrusive physiological sensors are being developed, this opens up a lot of possibilities for applications that employ physiological signals as their main emotion measure.

The finding that physiological signals are useful for inferring emotions is not surprising, as there have been many theories on emotion that have emphasized the role of physiological processes for emotions (James, 1884; Prinz, 2004). The low recognition performances for speech and facial expressions might be explained by the fact that our results come from data gathered in a non-social context. In a social context people might be more expressive and speech and facial expressions might be more valuable (Bucy & Bradley, 2004; Fridlund, 1991; Fischer et al., 2004). Moreover, when machines would be able to extract semantic information from speech, the speech signal would also become more valuable. Nevertheless, there are many applications of affective computing in non-social contexts and it seems that physiological signals are at least most valuable there.

Because physiological signals are valuable for machine emotion recognition, they might also be useful for humans trying to recognize emotions. In the light of the poor human emotion recognition performance we found in our studies, it is worthwhile to look at technologies that could assist humans in recognizing others' emotions (Picard, 2009). Based on the role of physiological signals for machine performance, it seems that presenting physiological signals as emotional information about others might benefit humans as well (Janssen et al., 2010). Further research is necessary to test this idea and answer questions surrounding what information is most useful, how such information should be presented to the user, and what the influence is of human learning on the information that someone can extract from physiological signals. Moreover, a more nuanced distinction between emotions might be necessary in communication applications. For instance, distinctions between happy and sad might not be as relevant as distinctions between frustration, anger, or insecurity (Scheirer et al., 2002).

Conclusion

A comparison of machine emotion recognition with human emotion recognition shows that machine emotion recognition is likely to be mature enough to be incorporated in applications. Improving the recognition further could potentially be done by incorporating semantic and context information, as humans also benefit from such information. A necessary next step will be to develop applications that employ these emotion recognition techniques and validate them in the real world. For this, the emotion recognition techniques have to be implemented in real-time, and unobtrusive sensors are necessary. Our results show that using physiological signals is very promising for this, especially now that there are more and more wireless unobtrusive sensor platforms being developed.

For long, computers have been seen as unempathic artifacts that we, nonetheless, treat like we treat other humans (Reeves & Nass, 1996). Incorporating machine emotion recognition into computers can help them better understand us and respond more appropriately to us. This is likely to improve both current human computer interaction and enable new applications. Therefore, now that our study shows that machine emotion recognition is quite far developed already, it becomes more and more important to identify what applications can benefit from such technology and what applications cannot benefit from it. Application domains for which machine emotion recognition can be useful include games, media and entertainment, (mediated) human communication, relaxation devices, transport, education, and augmented decision making, among others. Subsequently, it can be investigated what the benefits and costs are of incorporating affective technologies into these application areas. In the end, we foresee a future in which machines can be emotionally in touch with us, improving existing user experiences and opening up opportunities for a myriad of new technologies and tools that can dramatically improve our happiness, health, and well-being.

References

Ahn, S. J., Bailenson, J. N., Fox, J. A., & Jabon, M. E. (2010). Using Automated Facial Expression Analysis for Emotion and Behavior Prediction. In K. Doeveling, C. von Scheve, & E. A. Konijn (Eds.), *Handbook of emotions and mass media* (pp. 349–369). New York: Routledge.

Amft, O., & Lukowicz, P. (2009). From Backpacks to Smartphones: Past, Present, and Future of Wearable Computers. *IEEE Pervasive Computing*, *8*, 8–13.

Arnold, M. B. (1960). *Emotion and personality.* New York: Columbia University Press.

Ayadi, M. E., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, *44*, 572–587.

Bachorowski, J., & Owren, M. J. (2008). Vocal expressions of emotion. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (pp. 196–210). New York: Guilford Press.

Bailenson, J. N., Pontikakis, E. D., Mauss, I. B., Gross, J. J., Jabon, M. E., Hutcherson, C. A., et al. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International Journal of Human-Computer Studies*, *66*(5), 303–317.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*, 614–636.

Berntson, G. G., Bigger, J. T., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., et al. (1997, November). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623–648.

Berntson, G. G., Cacioppo, J. T., & Quickley, K. S. (1993). Cardiac Psychophysiology and Autonomic Space in Humans: Empirical Perspectives and Conceptual Implications. *Psychological Bulletin*, *114*, 296–322.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* New York: Oxford university press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Boersma, P., & Weenink, D. (Eds.). (2010, April). *Praat: Doing phonetics by computer.* Available from `http://www.praat.org/`

Boucsein, W. (1992). *Electrodermal activity.* New York: Plenum Press.

Bowyer, K. W., Chang, K., & Flynn, P. (2004). A survey of approaches to three-dimensional face recognition. In J. Kittler, M. Petrou, & M. Nixon (Eds.), *Proceedings of the 17th international conference on pattern recognition* (Vol. 1, pp. 358–361). Cambridge, UK.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control* (Vol. 16). Oakland, CA: Holden-day.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation 1: Defensive and appetitive reactions in picture processing. *Emotion*, *1*(3), 276–298.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment Manikin and the semantic differential. *Journal of Behavioral Therapy and Experimental Psychiatry*, *25*, 49–59.

Bucy, E. P., & Bradley, S. D. (2004). Presidential expressions and viewer emotion: Counterempathic responses to televised leader displays. *Social Science Information*, *43*(1), 59–94.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., et al. (2004, October). Analysis of emotion recognition using facial expressions, speech and multimodal information. In (pp. 205–211). State College, PA, USA: New York, NY, USA: ACM.

Butler, E. A., Wilhelm, F. H., & Gross, J. J. (2006). Respiratory sinus arrhythmia, emotion, and emotion regulation during social interaction. *Psychophysiology*, *43*, 612–622.

Cacioppo, J. T., Berntson, G. G., Klein, D. J., & Poehlmann, K. M. (1997). The psychophysiology of emotion accross the lifespan. *Annual Reviev of Gerontology and Geriatrics*, *17*, 27–74.

Cacioppo, J. T., & Tassinary, L. (1990). Inferring psychological significance from physiological signals. *American Psychologist*, *45*, 16–28.

Cacioppo, J. T., Tassinary, L., & Berntson, G. G. (2000). *Handbook of Psychophysiology.* Cambridge, MA: Cambridge University Press.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*, 18–37.

Cannon, W. B. (1927). The James-Lange theory of emotion: A critical examination and an alternative theory. *American Journal of Psychology*, *39*, 10–124.

Chanel, G., Kierkels, J. J. M., Soleymani, M., & Pun, T. (2009). Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, *67*(8), 607–627.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*, 27.

Chang, C.-Y., Tsai, J.-S., Wang, C.-J., & Chung, P.-C. (2009, April). Emotion recognition with consideration of facial expressions and physiological signals. In (pp. 278–283). Nashville, TN, USA: IEEE.

Chellapilla, K., Larson, K., Simard, P., & Czerwinski, M. (2005). Computers beat humans at single character recognition in reading based human interaction proofs (HIPs). In (pp. 21–29).

Chen, L. S., Tao, H., Huang, T. S., Miyasato, T., & Nakatsu, R. (1998, December). Emotion recognition from audiovisual information. In (Vol. Chapter 2:, pp. 83–88). Redondo Beach, CA, USA: Piscataway, NJ, USA: IEEE Signal Processing Society.

Chen, M., Gonzalez, S., Vasilakos, A., Cao, H., & Leung, V. C. M. (2011). Body Area Networks: A survey. *Mobile Networks and Applications*, *16*(2), 171–193.

Cohen, I., Sebe, N., Garg, A., Chen, L. S., & Huang, T. S. (2003). Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, *91*(12), 160–187.

Cottrell, G. W., & Metcalfe, J. (1991). EMPATH: Face, emotion, and gender recognition using holons. In *Advances in neural information processing systems* (Vol. 3 (Part IX, pp. 564–571). San Mateo, CA: Morgan Kaufmann.

Cowie, R. (2009, December). Perceiving emotion: towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3515–3525.

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, *40*(1-2), 5–32.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., et al. (2001). Emotion Recognition in humancomputer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain.* New York: G.P. Putnam.

Darwin, C. (1872). *The Expression of Emotions in Man and Animals.* Londen: Harper Collins.

Ekman, P. (1992). Facial expressions of emotion: New findings, new questions. *Psychological Science*, *3*, 34–38.

Ekman, P. (1999). Basic Emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion.* Sussex, U.K.: Harper Collins.

Ekman, P. (2009, December). Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3449–3451.

Ekman, P., & Friesen, W. V. (1971). Constants across culture in the face of emotion. *Journal of Personality and Social Psychology*, *17*, 124–129.

Ekman, P., & Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., & Friesen, W. V. (1986). A pan-cultural expression of emotion. *Motivation and Emotion*, *10*, 159–168.

Ekman, P., Levenson, R. W., & Friesen, W. V. (1983). Autonomic nervous system activity distinguishes between emotions. *Science*, *221*, 1208–1210.

Essa, I. A., & Pentland, A. P. (1995, June). Facial expression recognition using a dynamic model and motion energy. In (pp. 360–367). Cambridge, MA: IEEE Computer Society Press.

Essa, I. A., & Pentland, A. P. (1997). Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(7), 757–763.

Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: A survey. *Pattern Recognition*, *36*(1), 259–275.

Ferucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., et al. (2010). Building Watson: An overview of the DeepQA project. *AI Magazine*, *31*(3), 59–79.

Fischer, A. H., Manstead, A. S. R., & Zaalberg, R. (2004). Social influences on the emotion process. *European Review of Social Psychology*, *14*, 171–201.

Fodor, J. A. (1983). *Modularity of mind: An essay of faculty psychology.* Cambridge, MA: MIT Press.

Fridlund, A. J. (1991). Sociality of solitary smiling: Potentiation by an implicit audience. *Journal of Personality and Social Psychology*, *60*, 229–240.

Frijda, N. H. (1986). *The Emotions.* New York: Cambridge University Press.

Gelder, B. de. (2009, December). Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3475–3484.

Gelder, B. de, Van den Stock, J., Meeren, H. K. M., Sinke, C. B. A., Kret, M. E., & Tamietto, M. (2010). Standing up for the body. Recent progress in uncovering the networks involved in processing bodies and bodily expressions. *Neuroscience and Biobehavioral Reviews*, *34*, 513–527.

Gross, J. J., & John, O. P. (1997). Revealing feelings: Facets of emotional expressivity in self-reports, peer ratings, and behavior. *Journal of Personality and Social Psychology*, *72*, 435–448.

Grossman, P., & Taylor, E. W. (2007). Toward understanding respiratory sinus arrythmia: Relations to cardiac vagal tone, evolution and biobehavioral functions. *Biological Psychology*, *74*, 263–285.

Gunes, H., & Piccardi, M. (2007). Bi-modal emotion recogntion from expressive face and body gestures. *Journal of Network and Computer Applications*, *30*(4), 1334–1345.

Gunes, H., & Piccardi, M. (2009). Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics*, *39*(1), 64–84.

Haselager, W. F. G. (1997). *Cognitive science and folk psychology: The right frame of mind.* London: Sage Publications.

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, *6*(2), 156–166.

Herbert, B. M., Pollatos, O., Flor, H., Enck, P., & Schandry, R. (2010). Cardiac awareness and autonomic cardiac reactivity during emotional picture viewing and mental stress. *Psychophysiology*, *49*, 342–354.

Hoogen, W. M. van den, IJsselsteijn, W. A., & Kort, Y. A. W. de. (2008). Exploring Behavioral Expressions of Player Experience in Digital Games. In A. Nijholt & R. Poppe (Eds.), *Proceedings of the workshop on facial and bodily expression for control and adaptation of games ecag 2008* (pp. 11–19). Amsterdam, The Netherlands.

Hosseini, S. A., Khalilzadeh, M. A., & Changiz, S. (2010). Emotional stress recognition system for affective computing based on bio-signals. *Journal of Biological Systems*, *18*(1), 101–114.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). *A Practical Guide to Support Vector Classification Title.*

Hsu, F.-H. (2002). *Behind deep blue: Building the computer that defeated the world chess champion.* Princeton, NJ: Princeton University Press.

Ickes, W. J. (1997). *Empathic accuracy.* New York: Guilford Press.

Ickes, W. J., & Aronson, E. (2003). *Everyday mind reading: Understanding what other people think* *and feel.* New York: Prometheus Books.

Ickes, W. J., Gesn, P. R., & Graham, T. (2000). Gender differences in empathic accuracy: Differential ability or differential motivation? *Personal Relationships*, *7*(1), 95–109.

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

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37.

James, W. (1884). The Physical Basis of Emotion. *Psychological Review*, *1*, 516–529.

Janssen, J. H., Bailenson, J. N., IJsselsteijn, W. A., & Westerink, J. H. D. M. (2010). Intimate heartbeats: Opportunities for affective communication technology. *IEEE Transactions on Affective Computing*, *1*, 72–80.

Janssen, J. H., Van den Broek, E. L., & Westerink, J. H. D. M. (2011). Tune in to your emotions: A robust personalized affective music player. *User Modeling and User-Adaptive Interaction*, *In Press*.

Jung, K., Kim, K. I., & Jain, A. K. (2004). Text information extraction in images and video: A survey. *Pattern Recognition*, *37*(5), 977–997.

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

Kappas, A. (2010). Smile when you read this, whether you like it or not: Conceptual challanges to affect detection. *IEEE Transactions on Affective Computing*, *1*, 38–41.

Katsis, C. D., Katertsidis, N., Ganiatsas, G., & Fotiadis, D. I. (2008). Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach. *IEEE Transactions on Systems, Man, and CyberneticsPart A: Systems and Humans*, *38*(3), 502–512.

Kim, J., & André, E. (2006). Emotion Recognition Using Physiological and Speech Signal in Short-Term Observation. *Lecture Notes in Computer Science (Perception and Interactive Technologies)*, *4021*, 53–64.

Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*(12), 2067–2083.

Kim, J., André, E., Rehm, M., Vogt, T., & Wagner, J. (2005, September). Integrating information from speech and physiological signals to achieve emotional sensitivity. In (pp. 809–812). Lisboa, Portugal: L2F - Spoken Language Systems Laboratory.

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

Kleinsmith, A., Bianchi-Berthouze, N., & Steed, A. (2011). Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics*, in press–in press.

Knapp, R. B., Kim, J., & André, E. (2011). Physiological signals and their use in augmenting emotion recognition for humanmachine interaction. In P. Petta, C. Pelachaud, & R. Cowie (Eds.), *Emotion-oriented systems: The humaine handbook* (pp. 133–159). Berlin/Heidelberg, Germany: Springer-Verlag.

Kohavi, R., & John, G. (1997). Wrappers for feature subset selection. *Artificial intelligence*, *97*, 273–324.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*, 394–421.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, *52*, 372–385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1998). Motion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological Psychiatry*, *44*, 1248–1263.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: evaluative, facial, visceral, and behavioral responses. *Psychophysiology*, *30*, 261–273.

Lazarus, R. S. (1991). *Emotion and Adaptation.* New York: Oxford University Press.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review in Neuroscience*, *23*, 155–184.

Lemire, D. (2006). Streaming maximum-minimum filter using no more than three comparisons per element. *Nordic Journal of Computing*, *13*, 328–339.

Lemmens, P., Crompvoets, F. M. H., Brokken, D., Van den Eerenbeemd, J., & De Vries, J. J. G. (2009, March). A body-conforming tactile jacket to enrich movie viewing. In (pp. 7–12).

Levenson, R. W., Carstensen, L. L., Friesen, W. V., & Ekman, P. (1991). Emotion, physiology, and expression in old age. *Psychology and Aging*, *6*, 28–35.

Levenson, R. W., Ekman, P., Heider, K., & Friesen, W. V. (1992). Emotion and autonomic nervous system activity in the Minangkabau of west Sumatra. *Journal of Personality and Social Psychology*, *62*, 972–988.

Lewis, M., Haviland-Jones, J. M., & Barrett, L. F. (Eds.). (2008). *Handbook of Emotions.* New York: The Guilford Press.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. *The Journal of the Acoustical Society of America*, *34*(7), 922–927.

Lien, J. J.-J., Kanade, T., Cohn, J. F., & Li, C.-C. (2000). Detection, tracking, and classification of action units in facial expression. *Robotics and Autonomous Systems*, *31*(3), 131–146.

Lisetti, C. L., & Nasoz, F. (2004). Using Noninvasive Wearable Computers to Recogniza Human Emotions from Physiological Signals. *Journal of Applied Signal Processing*, *11*, 1672–1687.

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, *24*(6), 615–625.

Lorigo, L. M., & Govindaraju, V. (2006). Offline Arabic handwriting recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(5), 712–724.

Luengo, I., Navas, E., & Hernáez, I. (2010). Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on MultiMedia*, *12*(6), 490–501.

Lyons, W. (1986). *The disappearance of introspection.* Cambridge, MA: MIT Press.

Matsumoto, D., & Willingham, B. (2009). Spontaneous facial expressions of emotion in congenitally and non-congenitally blind individuals. *Journal of Personality and Social Psychology*, *96*, 1–10.

Mehrabian, A. (1970). A semantic space for nonverbal behavior. *Journal of Consulting and Clinical Psychology*, *35*, 248–257.

Meisel, W. S. (1972). *Computer-oriented approaches to pattern recognition* (Vol. 83). New York, NY, USA: Academic Press, Inc.

Mori, S., Suen, C. Y., & Yamamoto, K. (1992). Historical review of OCR research and development. *Proceedings of the IEEE*, *80*(7), 1029–1058.

Morrison, D., Wang, R., & Silva, L. C. D. (2007). Ensemble methods for spoken emotion recognition in call-centres. *Speech Communication*, *49*(2), 98–112.

Nwe, T. L., Foo, S. W., & Silva, L. C. D. (2003). Speech emotion recognition using hidden Markov models. *Speech Communication*, *41*(4), 603–623.

Nygaard, L., & Lunders, E. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & cognition*, *30*, 583–593.

Ochsner, K. N., Ray, R. R., Hughes, B., McRae, K., Cooper, J. C., Weber, J., et al. (2009). Bottom-Up and Top-Down Processes in Emotion Generation: Common and Distinct Neural Mechanisms. *Psychological Science*, *20*, 1322–1331.

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: Features and algorithms. *International Journal in Human-Computer Studies*, *59*(12), 157–183.

Pantic, M., Caridakis, G., André, E., Kim, J., Karpouzis, K., & Kollias, S. (2011). Multimodal emotion recognition from low-level cues. In P. Petta, C. Pelachaud, & R. Cowie (Eds.), *Emotion-oriented systems: The humaine handbook* (pp. 115–132). Berlin/Heidelberg, Germany: Springer-Verlag.

Pantic, M., & Patras, I. (2006). Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *36*(2), 433–449.

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

Peirce, C. S. (1933). *The collected papers of Charles Sanders Peirce* (C. Hartshorne, P. Weiss, & A. Burks, Eds.). Cambridge, MA: Harvard University Press.

Pentland, A. (2005). Socially aware computation and communication. *IEEE Computer*, *38*, 33–40.

Pentland, A. (2008). *Honest Signals: How they shape our world.* Cambridge, MA: MIT Press.

Petridis, S., & Pantic, M. (2011). Audiovisual discrimination between speech and laughter: Why and when visual information might help. *IEEE Transactions on Multimedia*, *13*(2), 216–234.

Petta, P., Pelachaud, C., & Cowie, R. (Eds.). (2011). *Emotion-oriented systems: The Humaine handbook.* Berlin/Heidelberg, Germany: Springer-Verlag.

Phelps, E. A. (2006). Emotion and cognition: Insights from studies of the human amygdala. *Annual Review of Psychology*, *57*, 27–53.

Phelps, E. A., O'Connor, K. J., Gatenby, J. C., Gore, J. C., Grillon, C., & Davis, M. (2001). Activation of the left amygdala to a cognitive representation of fear. *Nature Neuroscience*, *4*, 437–441.

Picard, R. W. (1997). *Affective Computing.* Cambridge, MA: MIT Press.

Picard, R. W. (2000). Toward agents that recognize emotion. *Vivek*, *13*(1), 3–13.

Picard, R. W. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies*, *59*, 55–64.

Picard, R. W. (2009). Future affective technology for autism and emotion communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3575–3584.

Picard, R. W., & Klein, J. (2002). Computers that recognise and respond to user emotion: theoretical and practical implications. *Interacting with Computers*, *14*, 141–169.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(10), 1175–1191.

Piferi, R. L., Kline, K. A., Younger, J., & Lawler, K. A. (2000). An alternative approach for achieving cardiovascular baseline: Viewing an aquatic video. *International Journal of Psychophysiology*, *37*, 207–217.

Porges, S. W., Saul, J. P., Stone, P. H., Van der Molen, M. W., Berntson, G. G., Bigger, J. T. J., et al. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, *34*, 623–647.

Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion.* New York: Oxford University Press.

Pylyshyn, Z. W. (1987). *The robot's dilemma.* Norwood: ABLEX Publishing Corporation.

Rani, P., Liu, C., Sarkar, N., & Vanman, E. (2006). An empirical study of machine learning techniques for affect recognition in human-robot interaction. *Pattern Analysis & Applications*, *9*(1), 58–69.

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* New York: Cambridge University Press.

Roth, G., & Dicke, U. (2005). Evolution of the brain and intelligence. *Trends in Cognitive Sciences*, *9*(5), 250–257.

Russell, J. A. (2003). Core Affect and the Psychological Construction of Emotion. *Psychological Review*, *110*, 145–172.

Russell, S., & Norvig, P. (1995). *Artificial Intelligence: A modern approach.* Upper Saddle River, NJ: Prentice Hall.

Ryle, G. (1949). *The concept of mind.* Chicago, IL: University of Chicago Press.

Sánchez, A., Ruiz, J. V., Moreno, A. B., Montemayor, A. S., Hernández, J., & Pantrigo, J. J. (2010). Differential optical flow applied to automatic facial expression recognition. *Neurocomputing*, *74*(8), 1272–1282.

Sarrafzadeh, A., Alexander, S., Dadgostar, F., Fan, C., & Bigdeli, A. (2008). How do you know that I don't understand? A look at the future of intelligent tutoring systems. *Computers in Human Behavior*, *24*, 1342–1363.

Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the user on purpose: A step toward building an affective computer. *Interacting with Computers*, *14*, 93–118.

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin*, *99*(2), 143–165.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication*, *40*(1-2), 227–256.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion Inferences from Vocal Expression Correlate Across Languages and Cultures. *Journal of Cross-Cultural Psychology*, *32*(1), 76–92.

Scherer, K. R., Bänziger, T., & Roesch, E. B. (Eds.). (2010). *Blueprint for affective computing: A sourcebook.* New York: Oxford University Press.

Scherer, K. R., & Zentner, M. R. (2001). Emotional effects of music: Production rules. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research.* New York: Oxford University Press.

Scherer, S., Schwenker, F., & Palm, G. (2009). Classifier fusion for emotion recognition from speech. In W. Minker, M. Weber, H. Hagras, V. Callagan, & A. D. Kameas (Eds.), *Advanced intelligent environments* (pp. 95–117). New York, NY, USA: Springer Science+Business Media, LLC.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, *53*, 1062–1087.

Silva, L. C. D., Miyasato, T., & Nakatsu, R. (1997, September). Facial emotion recognition using multi-modal information. In (Vol. 1, pp. 397–401). Singapore, Singapore: Piscataway, NJ, USA: IEEE Singapore Section / IEEE Signal Processing Society.

Sinha, R., & Parsons, O. A. (1996). Multivariate response patterning of fear. *Cognition and Emotion*, *10*(2), 173–198.

Sobol-Shikler, T., & Robinson, P. (2010). Classification of complex information: Inference of co-occurring affective states from their expressions in speech. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(7), 1284–1297.

Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, *27*(1), 24–27.

Tawari, A., & Trivedi, M. M. (2010). Speech emotion analysis: Exploring the role of context. *IEEE Transactions on MultiMedia*, *12*(6), 502–509.

Teasdale, J. D., Howard, R. J., Cox, S. G., Ha, Y., Brammer, M. J., Williams, S. C. R., et al. (1999). Functional MRI study of the cognitive generation of affect. *American Journal of Psychiatry*, *156*, 209–215.

Tian, Y.-L., Kanade, T., & Cohn, J. F. (2005). Facial expression analysis. In S. Z. Li & A. K. Jain (Eds.), *Handbook of face recognition* (pp. 247–275). New York, NY, USA: Springer Science+Business Media, Inc.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *49*, 433–460.

Van den Broek, E. L. (2010). Beyond biometrics. *Procedia Computer Science*, *1*, 2505–2513.

Van den Broek, E. L. (2011). Ubiquitous Empathy. *Personal and Ubiquitous Computing*.

Van den Broek, E. L., Janssen, J. H., Westerink, J. H. D. M., & Healey, J. A. (2009). Prerequisists for affective signal processing (ASP). In P. Encarnaca & A. Veloso (Eds.), *Biosignals 2009: Proceedings of the international conference on bio-inspired systems and signal processing* (pp. 426–433). Porto Portugal.

Van den Broek, E. L., Lisý, V., Janssen, J. H., Westerink, J. H. D. M., Schut, M. H., & Tuinenbreijer, K. (2010). Affective Man-Machine Interface: Unveiling human emotions through biosignals.

In A. Fred, J. Filipe, & H. Gamboa (Eds.), *Biomedical engineering systems and technologies: Biostec2009 selected revised papers* (Vol. 52, pp. 21–47). Berlin/Heidelberg, Germany: Springer.

Van den Broek, E. L., Schut, M. H., Westerink, J. H. D. M., & Tuinenbreijer, K. (2009). Unobtrusive Sensing of Emotions (USE). *Journal of Ambient Intelligence and Smart Environments*, *1*(3), 287–299.

Van den Broek, E. L., Van der Sluis, F., & Dijkstra, T. (2011). Telling the story and re-living the past: How speech analysis can reveal emotions in post-traumatic stress disorder (PTSD) patients. In J. H. D. M. Westerink, M. Krans, & M. Ouwerkerk (Eds.), *Sensing emotions: The impact of context on experience measurements* (Vol. 12, pp. [in press]–[in press]). Dordrecht, The Netherlands: Springer Science + Business Media B.V.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, *48*(9), 1162–1181.

Wagenaar, W. A. (1969). Note on the construction of digram-balanced Latin squares. *Psychological Bulletin*, *72*, 384–386.

Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia*, *10*(5), 936–946.

Westerink, J. H. D. M., De Vries, G., De Waele, S., Eerenbeemd, J., Van Boven, M., & Ouwerkerk, M. (2009). Emotion measurement platform for daily life situations. In A. Nijholt, J. Cohn, & M. Pantic (Eds.), *Proceedings of the international conference on affective computing and intelligent interaction* (pp. 217–223). Amsterdam, The Netherlands: IEEE.

Whang, M. (2008). The emotional computer adaptive to human emotion. In J. H. D. M. Westerink, M. Ouwerkerk, T. J. M. Overbeek, W. F. Pasveer, & B. De Ruyter (Eds.), *Probing experience: From assessment of user emotions and behaviour of development of products* (pp. 209–220). Dordrecht, The Netherlands: Springer.

Wilder, J. (1967). *Stimulus and response: The law of initial values.* Bristol, UK: Wright.

Wilhelm, F. H., & Grossman, P. (2010). Emotions beyond the laboratory: Theoretical fundaments, study design, and analytic strategies for advanced ambulatory assessment. *Biological Psychology*, *84*, 552–569.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* Amsterdam: Elsevier.

Wu, S., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, *53*(5), 768–785.

Xiao, R., Zhao, Q., Zhang, D., & Shi, P. (2011). Facial expression recognition on multiple manifolds.

Pattern Recognition, *44*(1), 107–116.

Yacoob, Y., & Davis, L. S. (2006). Recognizing human facial expressions from long image sequences using optical flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(6), 636–642.

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

Zaki, J., Bolger, N., & Ochsner, K. (2008). It Takes Two: The Interpersonal Nature of Empathic Accuracy. *Psychological Science*, *19*, 399–404.

Zaki, J., Bolger, N., & Ochsner, K. (2009). Unpacking the informational bases of empathic accuracy. *Emotion*, *9*, 478–487.

Zaki, J., Weber, J., Bolger, N., & Ochsner, K. (2009). The neural bases of empathic accuracy. *Proceedings of the National Academy of Sciences*, *106*(27), 11382–11387.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 39–58.

Zhai, J., & Barreto, A. (2006). Stress Detection in Computer Users Through Noninvasive Monitoring of Physiological Signals. *Biomedical Science Instrumentation*, *42*, 495–500.

Zhang, Y., & Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 699–714.