IGI PUBLISHING ITJ 5590
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, International Journal of Synthetic Emotions, Volume 1, Issue 1
edited by **Jordi Vallverdú © 2010, IGI Global**

Automatic, Dimensional and Continuous Emotion Recognition

Hatice Gunes, Imperial College London, UK

Maja Pantic, Imperial College London, UK and University of Twente, EEMCS, The Netherlands

ABSTRACT

Recognition and analysis of human emotions have attracted a lot of interest in the past two decades and have been researched extensively in neuroscience, psychology, cognitive sciences, and computer sciences. Most of the past research in machine analysis of human emotion has focused on recognition of prototypic expressions of six basic emotions based on data that has been posed on demand and acquired in laboratory settings. More recently, there has been a shift toward recognition of affective displays recorded in naturalistic settings as driven by real world applications. This shift in affective computing research is aimed toward subtle, continuous, and context-specific interpretations of affective displays recorded in real-world settings and toward combining multiple modalities for analysis and recognition of human emotion. Accordingly, this article explores recent advances in dimensional and continuous affect modelling, sensing, and automatic recognition from visual, audio, tactile, and brain-wave modalities.

Keywords: *Bodily Expression, Continuous Emotion Recognition, Dimensional Emotion Modelling, Emotional Acoustic and Bio-signals, Facial Expression, Multimodal Fusion*

INTRODUCTION

Human natural affective behaviour is multimodal, subtle and complex. In day-to-day interactions, people naturally communicate multimodally by means of language, vocal intonation, facial expression, hand gesture, head movement, body movement and posture, and possess a refined mechanism for understanding and interpreting information conveyed by these behavioural cues.

DOI: 10.4018/jse.2010101605

Despite the available range of cues and modalities in human-human interaction (HHI), the mainstream research on human emotion has mostly focused on facial and vocal expressions and their recognition in terms of seven discrete, basic emotion categories (neutral, happiness, sadness, surprise, fear, anger and disgust; Keltner & Ekman, 2000; Juslin & Scherer, 2005). In line with the aforementioned, most of the past research on automatic affect sensing and recognition has focused on recognition of facial and vocal expressions in terms of basic emotional states, and then based on data that

has been posed on demand or acquired in laboratory settings (Pantic & Rothkrantz, 2003; Gunes, Piccardi, & Pantic, 2008; Zeng, Pantic, Roisman, & Huang, 2009). Additionally, each modality—visual, auditory, and tactile—has been considered in isolation. However, a number of researchers have shown that in everyday interactions people exhibit non-basic, subtle and rather complex mental/affective states like thinking, embarrassment or depression (Baron-Cohen & Tead, 2003). Such subtle and complex affective states can be expressed via tens (or possibly hundreds) of anatomically possible facial expressions, bodily gestures or physiological signals. Accordingly, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such rich sources of information (Russell, 1980). Hence, a number of researchers advocate the use of dimensional description of human affect, where an affective state is characterized in terms of a small number of latent dimensions (e.g., Russell, 1980; Scherer, 2000; Scherer, Schorr, & Johnstone, 2001).

It is not surprising, therefore, that automatic affect sensing and recognition researchers have recently started exploring how to model, analyse and interpret the subtlety, complexity and continuity of affective behaviour in terms of latent dimensions, rather than in terms of a small number of discrete emotion categories.

A number of recent survey papers exist on automatic affect sensing and recognition (e.g., Gunes & Piccardi, 2008; Gunes et al., 2008; Zeng et al., 2009). However, none of those focus on dimensional affect analysis. This article, therefore, sets out to explore recent advances in human affect modelling, sensing, and automatic recognition from visual (i.e., facial and bodily expression), audio, tactile (i.e., heart rate, skin conductivity, thermal signals etc.) and brainwave (i.e., brain and scalp signals) modalities by providing an overview of theories of emotion (in particular the dimensional theories), expression and perception of emotions, data acquisition and annotation, and the current state-of-the-art in automatic sensing and recognition of emotional displays using a dimensional (rather than categorical) approach.

BACKGROUND RESEARCH

Emotions are researched in various scientific disciplines such as neuroscience, psychology, and linguistics. Development of automated affective multimodal systems depends significantly on the progress in the aforementioned sciences. Accordingly, we start our analysis by exploring the background in emotion theory, and human perception and recognition.

THEORIES OF EMOTION

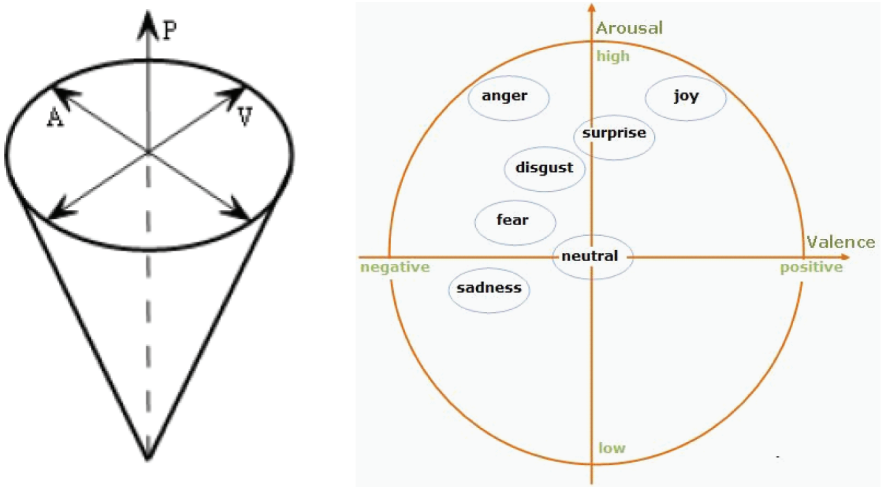
According to the research in psychology, three major approaches to emotion modelling can be distinguished (Grandjean, Sander, & Scherer, 2008): (1) categorical approach, (2) dimensional approach, and (3) appraisal-based approach.

The categorical approach is based on research on basic emotions, pioneered by Darwin (1998), interpreted by Tomkins (1962, 1963) and supported by findings of Ekman & his colleagues (1992, 1999). According to this approach there exist a small number of emotions that are basic, hard-wired in our brain, and recognized universally (e.g., Ekman & Friesen, 2003). Ekman and his colleagues conducted various experiments on human judgment of still photographs of deliberately displayed facial behaviour and concluded that six basic emotions can be recognized universally. These emotions are happiness, sadness, surprise, fear, anger and disgust (Ekman, 1982). Although psychologists have suggested a different number of such basic emotions, ranging from 2 to 18 categories (Ortony & Turner, 1990; Wierzbicka, 1992), there has been considerable agreement on the aforementioned six emotions. To date, Ekman's theory on universality and interpretation of affective nonverbal expressions in terms of basic emotion categories has been the most commonly adopted approach in research on automatic affect recognition.

On the other hand, however, a number of researchers in psychology argued that it is necessary to go beyond discrete emotions. Among various classification schemes, Baron-Cohen and his colleagues, for instance, have investigated cognitive mental states (e.g., agreement, concentrating, disagreement, thinking, reluctance, and interest) and their use in daily life. They did so via analysis of multiple asynchronous information sources such as facial actions, purposeful head gestures, and eye-gaze direction. They showed that cognitive mental states occur more often in everyday interactions than the basic emotions (Baron-Cohen & Tead, 2003). These states were also found relevant in representing problem-solving and decision-making processes in human-computer Interaction (HCI) context and have been used by a number of researchers, though based on deliberately displayed behaviour rather than in natural scenarios (e.g., El Kaliouby & Robinson, 2005).

According to the dimensional approach, affective states are not independent from one another; rather, they are related to one another in a systematic manner. In this approach, majority of affect variability is covered by three dimensions: valence, arousal, and potency (dominance) (Davitz, 1964; Mehrabian & Russell, 1974; Osgood, Suci, & Tannenbaum, 1957). The valence dimension refers to how positive or negative the emotion is, and ranges from unpleasant feelings to pleasant feelings of happiness. The arousal dimension refers to how excited or apathetic the emotion is, and it ranges from sleepiness or boredom to frantic excitement. The power dimension refers to the degree of power or sense of control over the emotion. Taking into account the aforementioned, a reasonable space of emotion can be modelled as illustrated in Figure 1a. Russell (1980) introduced a circular configuration called Circumflex of Affect (see Figure 1b) and proposed that each basic emotion represents a bipolar entity being a part of the same emotional continuum. The proposed polars are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). As illustrated in Figure 1b, the proposed emotional space consists of four quadrants: low arousal positive, high arousal

Figure 1. Illustration of a) three dimensions of emotion space (V-valence, A-arousal, P-power), and b) distribution of the seven emotions in arousal-valance (A-V) space. Images adapted from (Jin & Wang, 2005) and (Breazeal, 2003), respectively.



(a) (b)

positive, low arousal negative, and high arousal negative. In this way, as argued by Russell, it is possible to characterize all emotions by their valence and arousal, and different emotional labels could be plotted at various positions on this two-dimensional plane.

However, each approach, categorical or dimensional, has its advantages and disadvantages. In the categorical approach, where each affective display is classified into a single category, complex mental/affective state or blended emotions may be too difficult to handle (Yu, Aoki, & Woodruff, 2004). Instead, in dimensional approach, observers can indicate their impression of each stimulus on several continuous scales. Despite exhibiting such advantages, dimensional approach has received a number of criticisms. Firstly, the usefulness of these approaches has been challenged by discrete emotions theorists, such as Silvan Tomkins, Paul Ekman, and Carroll Izard, who argued that the reduction of emotion space to two or three dimensions is extreme and resulting in loss of information. Secondly, while some basic emotions proposed by Ekman, such as happiness or sadness, seem to fit well in the dimensional space, some basic emotions become indistinguishable (e.g., fear and anger), and some emotions may lie outside the space (e.g., surprise). It also remains unclear how to determine the position of other affect-related states such as confusion. Note, however, that arousal and valence are not claimed to be the only dimensions or to be sufficient to differentiate equally between all emotions. Nonetheless, they have proven to be useful in several domains (e.g., affective content analysis as reported by Yang, Lin, Su, & Chen, 2007).

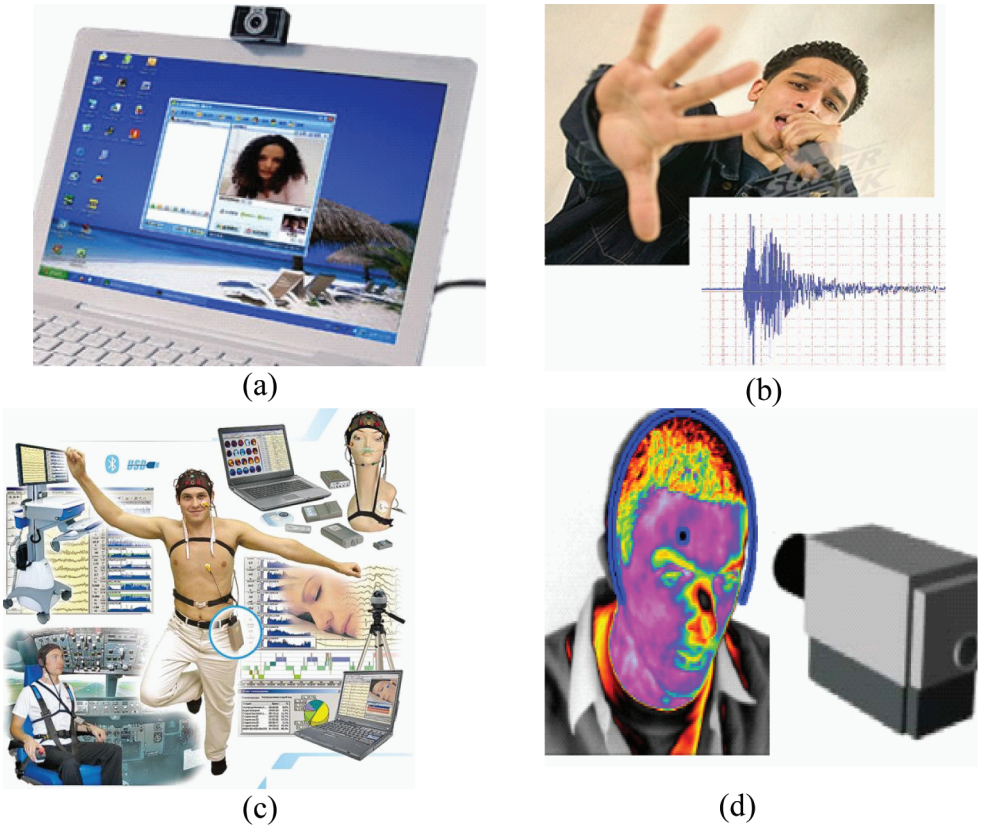
Scherer and colleagues introduced another set of psychological models, referred to as componential models of emotion, which are based on appraisal theory (Scherer et al., 2001). The appraisal-based approach, which can also be seen as extension to the dimensional approach, claims that emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world. This approach views emotions through changes in all relevant components including cognition, motivation, physiological reactions, motor expressions, and feelings. The advantage of componential models is that they do not limit emotional states to a fixed number of discrete categories or to a few basic dimensions. Instead, they focus on the variability of different emotional states, as produced by different types of appraisal patterns. Emotion is described through a set of stimulus evaluation checks, including the novelty, intrinsic pleasantness, goal-based significance, coping potential, and compatibility with standards. Therefore, differentiating between various emotions and modelling individual differences become possible. How to use the appraisal-based approach for automatic emotion recognition remains an open research question due to the fact that this approach requires complex, multicomponential and sophisticated measurements of change.

Even with over a century of research, all of the aforementioned issues, and in particular the issue of which psychological model of emotion is more appropriate for which context, still remain under discussion. For further details on different approaches to modelling human emotions and their relative advantages and disadvantages, the reader is referred to the works by Scherer (2000) and Grandjean et al. (2008).

EXPRESSION AND PERCEPTION OF EMOTIONS

Emotional information is conveyed by a broad range of multimodal cues, including speech and language, gesture and head movement, body movement and posture, vocal intonation and facial expression, and so forth. Herewith, we provide a summary of the findings from research on emotion communication by means of facial and bodily expression, speech and nonverbal vocalizations, bio-potential signals (physiological signals, brain waves and thermal signals). Figure 2 illustrates examples of sensors used for acquiring affective data from these cues and modalities.

Figure 2. Examples of sensors used in multimodal affective data acquisition: (a) camera for visible imagery, (b) microphone(s) for audio recording, (c) various sensors for bio-potential signal recording and (d) infrared camera for thermal imagery



FACIAL EXPRESSION

Ekman and his colleagues conducted various experiments of human judgment on still photographs of deliberately displayed facial behaviour and concluded that six basic emotions can be recognized universally: happiness, sadness, surprise, fear, anger and disgust. Several other emotions and many combinations of emotions have been studied as well but it remains unconfirmed whether they are universally distinguishable. Although prototypic expressions of basic emotions like happiness, surprise, and fear are natural, they occur infrequently in daily life and provide an incomplete description of facial behaviour. To capture the subtlety of human facial behaviour, Ekman and Friesen developed the Facial Action Coding System (FACS) for coding fine-grained changes in the face (Ekman & Friesen, 1978; Ekman, Friesen, & Hager, 2002). FACS is based on the enumeration of all *facial action units*, which are related to facial muscle actions, causing changes in the facial appearance. In addition to this, Friesen and Ekman (1984) developed Emotion FACS (EMFACS) as a method for using FACS to score only the facial actions that might be relevant to detecting emotions.

As proposed by a number of researchers (e.g., Plutchik, 1984; Russell, 1997), different facial expressions could also be mapped to various positions on the two-dimensional plane of

arousal-valence. This is illustrated in Figure 1b, where a series of facial expression photos was mapped onto Russell's (1997) arousal-valence dimensions (Breazeal, 2003).

To date, however, Ekman's theory of basic emotions and the FACS are the most commonly used schemes in vision-based systems attempting to recognize facial expressions and analyze human affective behaviour (Pantic & Rothkrantz, 2003; Zeng et al., 2009).

BODILY EXPRESSION

Researchers in social psychology and human development have long emphasized the fact that emotional states are expressed through body movement (Argyle, 1975; Darwin, 1998; Hadjikhani & De Gelder, 2003). However, compared to research on facial expression, the expressive information body gestures carry has not been adequately explored yet.

The main focus has been that of mapping bodily expression onto discrete emotion categories. Darwin (1998) was the first to describe in detail the bodily expressions associated with emotions in animals and humans and proposed several principles underlying the organization of these expressions. Following Darwin's early work, there have been a number of studies on human body postures communicating emotions (e.g., Argyle, 1975). Coulson presented experimental results on attribution of six emotions (anger, disgust, fear, happiness, sadness and surprise) to static body postures by using computer-generated figures (Coulson, 2004). He found out that in general, human recognition of emotion from posture is comparable to recognition from the voice, and some postures are recognized as effectively as facial expressions.

Van den Stock, Righart, and De Gelder (2007) also presented a study investigating emotional body postures (happiness, sadness, surprise, fear, disgust and anger) and how they are perceived. Results indicate good recognition of all emotions, with angry and fearful bodily expressions less accurately recognized compared to, for example, bodily expressions of sadness.

Behavioural studies have shown that posture can communicate affective dimensions as well as discrete emotion categories. Kleinsmith, Ravindra De Silva, and Bianchi-Berthouze (2005) identified that scaling, arousal, valence, and action tendency were the affective dimensions used by human observers when discriminating between postures. They reported that low-level posture features such as orientation (e.g., orientation of shoulder axis) and distance (e.g., distance between left elbow and left shoulder) could effectively discriminate between the affective dimensions.

In general, dimensional models are considered important in affect sensing as a single label may not reflect the complexity of the affective state conveyed by a body posture or gesture. It is also worth noting that Ekman and Friesen (1967) considered expressing discrete emotion categories via face, and communicating dimensions of affect via body as more plausible. However, communication of emotions by bodily movement and expressions is still a relatively unexplored and unresolved area in psychology, and further research is needed in order to obtain a better insight on how they contribute to the perception and recognition of various affective states both in terms of categories and A-V dimensions.

AUDIO

Speech conveys affective information through explicit (linguistic) messages, and implicit (paralinguistic) messages that reflect the way the words are spoken. If we consider the verbal part (linguistic message) only, without regarding the manner in which it was spoken (paralinguistic message), we might miss important aspects of the pertinent utterance and even misunderstand the spoken message by not attending to the nonverbal aspect of the speech. However, findings in basic research indicate that spoken messages are rather unreliable means to analyze and predict human (affective) behaviour (Ambady &

Rosenthal, 1992). Anticipating a person's word choice and the associated intent is very difficult: even in highly constrained situations, different people choose different words to express exactly the same thing. Yet, some information about the speaker's affective state can be inferred directly from the surface features of words, which were summarized in some affective word dictionaries and lexical affinity (e.g., Whissell, 1989). The rest of affective information lies below the text surface and can only be detected when the semantic context (e.g., discourse information) is taken into account. The association between linguistic content and emotion is language-dependent and generalizing from one language to another is very difficult to achieve (Ortony & Turner, 1990).

When it comes to implicit, paralinguistic messages that convey affective information, the research in psychology and psycholinguistics provides an immense body of results on acoustic and prosodic features which can be used to encode affective states of a speaker. For a comprehensive overview of the past research in the field, readers are referred to Juslin and Scherer (2005). The prosodic features which seem to be reliable indicators of the basic emotions are the continuous acoustic measures, particularly pitch-related measures (range, mean, median, and variability), intensity and duration. For a comprehensive summary of acoustic cues related to vocal expressions of basic emotions, readers are referred to Cowie et al. (2001). However, basic researchers have not identified an optimal set of voice cues that reliably discriminate among emotions. Nonetheless, listeners seem to be accurate in decoding some basic emotions from prosody (Juslin & Scherer, 2005) as well as some non-basic affective states such as distress, anxiety, boredom, and sexual interest from non-linguistic vocalizations like laughs, cries, and yawns (Russell & Fernández-Dols, 1997).

There have also been a number of works focusing on how to map audio expression to dimensional models. Cowie et al. used valence-activation space, which is similar to the A-V space, to model and assess emotions from speech (Cowie, Douglas-Cowie, Savvidou, McMahon, Sawey, & Schroder, 2000; Cowie et al., 2001). Scherer and his colleagues have also proposed how to judge emotion effects on vocal expression, using the appraisal-based theory (Grandjean et al., 2008; Scherer, 2000).

BIO-POTENTIAL SIGNALS

Numerous findings in psychophysiology suggest that the activation of the autonomic nervous system changes when emotions are elicited (Levenson, 1988).

While the visual modality including facial expressions and body gestures provides a visible proof of affective arousal, bio-signals such as electroencephalography (EEG) and functional near-infrared spectroscopy (fNIRS) provide an invisible proof of affective arousal (Savran et al., 2006). The signals commonly referred to as physiological or bio-signals (Changchun, Rani, & Sarkar, 2005; Savran et al., 2006; Takahashi, 2004) and used in affect sensing research field to identify emotions can be listed and described as follows.

- Galvanic Skin Response (GSR) provides a measurement of the of skin conductance (SC). SC increases linearly with a person's level of overall arousal or stress (Chanel, Kronegg, Grandjean, & Pun, 2007).
- Electromyography (EMG) measures the muscle activity or frequency of muscle tension, and has been shown to correlate with negatively valenced emotions (Haag, Goronzy, Schaich, & Williams, 2004, Nakasone, Prendinger, & Ishizuka, 2005).
- Blood Volume Pulse (BVP) is an indicator of blood flow. Since each heart beat (or pulse) presses blood through the vessels, BVP can also be used to calculate heart rate and inter-beat intervals. Heart rate increases with negatively valenced emotions, such as anxiety or fear.
- Skin temperature (ST) describes the temperature as measured on the surface of the skin.

- Electrocardiogram (ECG) signal measures contractile activity of the heart. This can be recorded either directly on the surface of the chest or alternatively on the limbs (more sensitive to artefacts). It can be used to measure heart rate and inter-beat intervals to determine the heart rate variability (HRV). A low HRV can indicate a state of relaxation, whereas an increased HRV can indicate a potential state of mental stress or frustration.
- Respiration rate (R) measures how deep and fast a person is breathing. Slow and deep breathing indicates a relaxed resting state while irregular rhythm, quick variations, and cessation of respiration corresponds to more aroused emotions like anger or fear (Chanel et al., 2007; Haag et al., 2004).

There is evidence suggesting that measurements recorded over various parts of the brain including the amygdala enable observation of the emotions felt (Pun et al., 2006). For instance, approach or withdrawal response to a stimulus is known to be linked to the activation of the left or right frontal cortex, respectively.

As stated by Arroyo-Palacios and Romano (2008) physiological or bio-signals offer great possibilities for automatic affect recognition. However, exploiting their full potential has been impossible to date due to a lack of consensus among psychologists about the nature, theories, models, and specificity of physiological patterns for each emotion-space dimension. Needless to say, establishing standardization on key areas such as stimulus for the identification of physiological patterns, physiological measures, features to analyze, and the emotional model to be used will greatly advance the state-of-the-art in this field (Arroyo-Palacios & Romano, 2008).

THERMAL SIGNALS

A number of studies in neuropsychology, physiology, and behaviour analysis suggest that there exists a correlation between mammals' core body temperature and their affective states. Nakayama, Goto, Kuraoka, & Nakamura (2005) conducted experiments by monitoring the facial temperature change of monkeys under stressful and threatening conditions. Their study revealed that a decrease in nasal skin temperature is relevant to a change from neutral to negative affective state. Vianna and Carrive (2005) conducted another independent experiment by monitoring the temperature changes in rats when they were experiencing fearful situations. They observed that the temperature increased in certain body parts (i.e., eyes, head and back), while in other body parts (i.e., tail and paws) the temperature dropped simultaneously.

Other studies also exist indicating that contraction or expansion of the facial/bodily muscles of humans causes fluctuations in the blood flow rate (e.g., Khan, Ingleby, & Ward, 2006, Khan, Ward, & Ingleby, 2006, Khan, Ward, & Ingleby, 2009; Tsiamyrtzis, Dowdall, Shastri, Pavlidis, Frank, & Ekman, 2007). This muscular activity results in a change in the volume of blood flow under the surface of the human facial and/or bodily skin. Thus, tensed or contracted muscles (e.g., in anger or stress) result in higher skin temperature.

Unlike other bio-physiological sensing, the use of infrared thermal camera does not rely on contact with the human body. Thus, non-invasive detection of any change in facial and/or bodily thermal features relevant to detecting, extracting, and interpreting human affective states is feasible. For instance, Pavlidis, Levine, and Baukol (2001) and Tsiamyrtzis et al. (2007) have shown that there is a correlation between increased blood perfusion in the orbital muscles and anxiety and stress levels of humans. Similarly, Puri, Olson, Pavlidis, Levine, and Starren (2005) reported that users' stress level was correlated with increased blood flow in the frontal vessels of forehead causing dissipation of convective heat.

A generic model for estimating the relationship between fluctuations in blood flow, skin temperature, and facial/bodily muscle activity is not yet available. Such a model could enhance

our understanding of the relationship between affective dimensions and the facial/bodily thermal and physiological reactions.

POSED VS. SPONTANEOUS EXPRESSIONS

Most of the studies supporting the universality of emotional expressions are based on experiments related to deliberate/posed expressions. Studies reveal that humans recognise both deliberate/posed and involuntary/spontaneous emotional expressions equally accurately. However, deliberate expressions are significantly different from spontaneous expressions. Deliberate facial behaviour is mediated by separate motor pathways and differences between natural and deliberate facial actions may be significant. Schmidt and Cohn (2001) reported that an important visual cue signalling a smile as being deliberate or spontaneous is the temporal evolvement of the smile. Extensive research has been further conducted by Cohn and colleagues in order to identify temporal and morphological differences between deliberately and spontaneously displayed facial affective behaviour (Ambadar, Schooler, & Cohn, 2005).

In daily interactions, a particular bodily expression is most likely to be accompanied by a congruent facial expression being governed by a single emotional state. Darwin argued that because our bodily actions are easier to control on command than our facial actions, the information conveyed by body movements should be less significant than that conveyed by the face, at least when it comes to discerning spontaneous from posed behaviour. Ekman, however, argued that people do not bother to censor their body movements in daily life; therefore, the body would be the more reliable source of information (Ekman, 2003). This is also in agreement with recent findings in research in nonverbal behaviour and communication, which state that truthful and deceptive behaviour differ from each other in lack of head movement (Buller, Burgoon, White, & Ebesu, 1994) and lack of illustrating gestures which accompany speech (DePaulo, 2003) in the case of deceptive behaviour.

Compared to visible channels of face and body, the advantage of using bio-signals for recognizing affective states is the fact that physiological recordings cannot be easily faked or suppressed, and can provide direct information about the user's affective state.

However, people express and communicate emotions multimodally. Hence, more research efforts and studies on posed vs. spontaneous expressions in a multicue and multimodal context are needed if we are to obtain a better understanding of the natural communication of emotions in HHI to be later used in HCI.

DATA ACQUISITION

Recordings of affective behaviour may be those of posed behaviour (i.e., produced by the subject upon request), induced behaviour (i.e., occurring in a controlled setting designed to elicit an affective reaction such as when watching movies), or spontaneous behaviour (i.e., occurring in real-life settings such as interviews or interactions between humans or between humans and machines) (Banziger & Scherer, 2007).

The easiest way to create a database of acted affective displays is by having an experimenter direct and control the recorded displays. Depending on which modalities are recorded, a number of sensors can be used: cameras for face and body expressions, microphones for recording audio signals, a motion capture systems to record 3D affective postures/gestures, and so forth. (see Figure 2). When acquiring spontaneous affective multimodal data, the subjects may be recorded without their knowledge while they are stimulated with some emotionally-rich stimulus (e.g., Zuckerman, Larrance, Hall, DeFrank, & Rosenthal, 1979). Due to the ethical issues, making recordings without subjects' knowledge is strongly discouraged and the current trend is to record spontaneous data in more constrained conditions such as an interview settings, where subjects are still aware of placement of cameras

and their locations (e.g., Littlewort, Bartlett, & Lee, 2007; Pantic & Bartlett, 2007).

3D affective body postures or gestures can alternatively be recorded by utilizing motion capture systems (e.g., Kleinsmith & Bianchi-Berthouze, 2007). In such scenarios, the actor is dressed in a suit with a number of markers on the joints and body segments, while each gesture is captured by a number of cameras and represented by consecutive frames describing the position of the markers in the 3D space.

Recording physiological and bio-potential signals is a bit more complicated compared to the aforementioned recordings. In the brain-computer interface (BCI) or bio-potential signal research context, the subject being recorded usually wears headphones, a headband or a cap on which electrodes are mounted, a clip sensor, and/or touch type electrodes. The subject is then stimulated with emotionally-evocative images/videos/sounds. The variation of the skin conductance at the region of interest is then measured (Takahashi, 2004). Hence, the bio-potential affect data acquisition is *induced* and, due to its invasive nature, the experimental settings provided do not encourage spontaneity.

Creation and annotation of affect databases from face and body displays has been reviewed by Gunes and Piccardi (2006). Various visual, audio and audio-visual databases have been reviewed by Zeng et al. (2009). The existing databases where emotion is labelled continuously and data were made publicly available for research purposes are listed in Table 1. Overall, very few of the existing multimodal affect databases contain spontaneous data. Although there is a recent attempt to collect spontaneous facial expression data in real-life settings (in

Table 1. Representative databases created for dimensional affect recognition.

Database	The Montreal Affective Voices Database	SAL database	The Vera am Mittag speech database
Reference	Belin, Fillion-Bilodeau, and Gosselin, 2008	Douglas-Cowie et al., 2007	Grimm, Kroschel, and Narayanan, 2008
Data Type	Posed	induced	spontaneous
Modalities	Emotional speech	Audiovisual: facial expressions, emotional speech	Audiovisual: facial and bodily expressions, emotional speech
Subjects	5 male and 5 female actors	2 male and 2 female subjects interacting with an artificial listener	various participants in the show
Categorical Annotation	anger, disgust, sadness, fear, pain, happiness, pleasure, surprise, neutral	not applicable	not applicable
Dimensional Annotation	intensity of valence, intensity of arousal, and intensity of each discrete emotion category	intensity of arousal, and intensity of valence	continuous annotation for valence, activation, and dominance.
Annotators	30 observers	4 Feeltrace coders	17 observers
Content	90 emotionally-coloured pronunciations of the word 'ah'	Humans interacting with a Sensitive Artificial Listener (SAL) in a Wizard-of-Oz scenario	12 hours of audio-visual recordings of German TV talk show "Vera am Mittag", segmented into dialogue acts and utterances
Public Availability	Yes	yes	yes
Online Provision	No	no	no

the context of autism disorder; El Kaliouby & Teeters, 2007), such an attempt is lacking when it comes to multimodal human-affect data. As already mentioned above, acquiring data in fully unconstrained environments with multiple sensors involves ethical and privacy concerns together with numerous technical difficulties (placement of sensors, controlling the environmental conditions such as noise, illumination, occlusions; consistency, repeatability, etc.). This impedes significantly the progress in this direction.

DATA ANNOTATION

In general, annotation of the data, both for posed and spontaneous data, is usually done separately for each channel assuming independency between the channels.

In general, for databases containing audio data, the annotation tool FeelTrace is commonly used. Feeltrace allows observers to listen to affective behaviour recordings (and watch them in the case of audio-visual recordings) and move their cursor within a 2D emotional space to rate their impression about the emotional state of the subject (Cowie et al., 2000). Emotion classes based on Feeltrace can be described as follows: positive activation, positive evaluation; positive activation, negative evaluation; negative activation, negative evaluation; negative activation, positive evaluation; and neutral (close to the centre of the 2D emotional space). For instance, for the Sensitive Artificial Listener (SAL) database (see Douglas-Cowie et al., 2007; Table 1), 4 observers provided continuous annotations with respect to valence and activation dimensions, using the FeelTrace annotation tool.

In general, when annotating or labelling affective behaviour from facial displays, six basic emotion categories and the Facial Action Coding System (FACS) are used. There exist very few studies focusing on labelling facial expressions using the dimensional approaches. For instance, Breazeal (2003) mapped a series of facial expression photos onto Russell's A-V emotion space (see Figure 1b) and used this to model a robot's interpretation of facial expressions. Shin (2007) asked human observers to rate static facial expression images in terms of A-V dimensions on a nine-point scale. The images were labelled with a rating averaged over all observers. As described previously, the FeelTrace annotation tool is often used to annotate audio and audio-visual recordings (e.g., in the case of the SAL database).

When it comes to annotating body gestures, there is not one common annotation scheme that has been adopted by all research groups. Kleinsmith and Bianchi-Berthouze (2007) reported results for five observers that were asked to rate static body postures on a seven-point Likert scale in terms of four affective dimensions: valence (pleasure), arousal (alertness), potency (control), and avoidance (avoid/attend to). Postures that received an average observer rating of < 3.8 were labelled as low intensity postures. Postures that received an average rating between 3.8 and 4.2 were labelled as neutral intensity postures. Finally, postures that received an average rating of > 4.2 were labelled as high intensity postures.

Using the categorical and dimensional models simultaneously enables analysis of mapping between categorical and dimensional spaces. The Montreal Affective Voices Database (Belin et al., 2008), for instance, includes 10 ratings for each data sample: perceived valence (from extremely negative to extremely positive), perceived arousal (from not aroused to extremely aroused), and perceived intensity of eight targeted affective states: happiness, sadness, fear, anger, surprise, disgust, pleasure, and pain (e.g., from not angry to extremely angry). Jin and Wang (2005) analyzed emotions in spoken Chinese and reported that joy and anger are commonly associated with similar high level of arousal while surprise, disgust, fear, neutral, and sadness are commonly associated with lower levels of arousal. As far as valence is concerned, joy was commonly associated with high levels of valence. Differences in the ratings in terms of the arousal dimension were reported to be smaller than those reported for the valence dimension.

Annotating brain-wave, thermal, and other signals in terms of affective states is not a straightforward process and it is inherently different compared to visual or audio recordings. For bio-potential signal annotation, the level of valence and arousal is usually extracted from the subjects' responses (Kulic & Croft, 2007; Pun, Alecu, Chanel, Kronegg, & Voloshynovskiy, 2006). This is mainly due to the fact that feelings induced by an image can be very different from subject to subject. Self-assessment of valence and arousal is therefore a preferred way of labelling the data (Chanel et al., 2005). The subjects are generally asked to rate their response to the stimuli in terms of intensity of few affect categories (Kulic & Croft, 2007). When a dimensional approach is used, intensity scores such as *low*, *medium* and *high* are usually scored for arousal and valence dimensions (Kulic & Croft, 2007).

Overall, researchers seem to use different levels of intensity when adopting a dimensional affect approach. Shin (2007) asked the observers to rate static facial expression images in terms of A-V using a ten-point Likert scale (0-very positive, 9-very negative), (0-low arousal, 9-high arousal). Yang et al. (2007) use a range between -1.0 and 1.0, divided into 11 levels, for annotation of emotions in the A-V space. The final annotation is then calculated as the mean of the A-V values of all observers.

Obtaining high inter-observer agreement is one of the main challenges in affective data annotation, especially when dimensional approach is adopted. Yang et al. (2007) report that mapping emotions onto the A-V space confuses the subjects. For instance, the first quadrant (high arousal, high valence) contains emotions such as excited, happy, and pleased, which are different in nature. In addition, the Feeltrace representation is criticized for not being intuitive, and raters seem to need special training to use such a dimensional labelling system (Zeng et al., 2009). A hybrid coding scheme combining both dimensional and categorical descriptions, similar to that of Zhang, Tian, Jiang, Huang, and Gao (2008), or a hierarchical scheme where the first level focuses on intensity (high, medium, low), and the second level focuses on emotions with high (happy, fear, anger), medium (happy, neutral, sad) and low (sad, neutral) arousal (Xu, Jin, Luo, & Duan, 2008) could potentially ease naïve observer's annotation task. Development of an easy to use, unambiguous and intuitive annotation scheme remains, however, an important challenge.

Another major challenge in affect data annotation is the fact that there is no coding scheme that is agreed upon and used by all researchers in the field and that can accommodate all possible communicative cues and modalities including facial and bodily expressions, vocal intonation and vocalization (e.g., laughter), bio-potential signals, etc. Addressing the aforementioned issues is necessary if we are to advance the state-of-the-art in dimensional affect sensing and recognition by making the research material comparable and easy to use.

AFFECT RECOGNITION

A typical approach to affect recognition is to categorize input samples into a number of emotion classes and apply standard pattern recognition procedures to train a classifier (Yang et al., 2007). This approach proved reasonably successful for categorical emotion recognition (Gunes et al., 2008; Pantic & Rothkrantz, 2003; Zeng et al., 2009). However, is this approach suitable when it comes to dimensional emotion recognition? We attempt to find answers to this question by examining the problem domain and surveying the state of the art in the field.

PROBLEM DOMAIN

Affect recognition is context dependent (sensitive to who the subject is, where she is, what her current task is, and when the observed behaviour has been shown; Pantic, Nijholt, & Petland, 2008). It must be carried out differently

in the case of acted behaviour than in the case of spontaneous behaviour (see the previous section of this article), and both configuration and temporal analysis of the observed behaviour are of importance for its interpretation (Ambadar et al., 2005). Except of these issues, which are typical for any human behaviour interpretation, and have been discussed in various papers (e.g., Pantic & Bartlett, 2007; Pantic et al., 2008; Vinciarelli, Pantic, & Bourland, 2009), there are a number of additional issues which need to be taken into account when applying a dimensional approach to emotion recognition. These include reliability of the ground-truth, determining duration of emotions for automatic analysis, determining the baseline, dimensionality reduction, modelling intensity of emotions, high inter-subject variation, defining optimal fusion of cues/modalities, and identifying appropriate classification methods and evaluation measures.

RELIABILITY OF THE GROUND-TRUTH

Achieving inter-observer agreement is one of the most challenging issues in dimension-based affect modelling and analysis. To date, researchers have mostly chosen to use self-assessments (e.g., Pun et al., 2006) or the mean (within a predefined range of values) of the observers' ratings (e.g., Kleinsmith & Bianchi-Berthouze, 2007). Chanel et al. (2005) report that although it is difficult to self-assess arousal, using classes generated from self-assessment of emotions facilitate greater accuracy in recognition. This finding results from a study on automatic analysis of physiological signals in terms of A-V emotion space (Chanel et al., 2005). It remains unclear whether the same holds independently of the utilised modalities and cues. Modelling inter-observer agreement levels within automatic affect analyzers and finding which signals better correlate with self assessments and which ones better correlate with independent observer assessments remain unexplored.

DURATION OF EMOTIONS

Determining the length of the temporal window for automatic affect analysis depends in principle on the modality and the target emotion. Levenson (1988) suggests that overall duration of emotions approximately falls between 0.5 and 4 seconds. He points out that, when measuring at wrong times, the emotion might be missed or multiple different emotions might be covered when too long periods are measured. For instance, when measuring bio-signals, for surprise the latency of onset can be very short, while for anger it may be rather long. Overall, the existing literature does not provide a unique answer regarding the window size to be used to achieve optimal affect recognition. Also, there is no consensus on how the efficiency of a choice should be evaluated. Current affect recognizers employ various window sizes depending on the modality, e.g., 2-6 seconds for speech, 3-15 seconds for bio-signals (Kim, 2007).

EMOTION INTENSITY

In dimensional emotion recognition the intensity of an emotion is encoded in the level of arousal (Kulic & Croft, 2007). Different emotions that have a similar level of valence can only be discriminated by their level of arousal. For instance, at a neutral valence level, low arousal represents *calmness* while high arousal represents *excitement*. Intensity is usually measured by modelling it with discrete levels such as neutral, low and high (e.g., Kleinsmith & Bianchi-Berthouze, 2007; Kulic & Croft, 2007; Wollmer et al., 2008). Separate models are then built to discriminate between pairs of affective dimension levels, for instance, low vs. high, low vs. neutral, etc. (Kleinsmith & Bianchi-Berthouze, 2007). Measuring the intensity of shown emotion appears to be modality dependent. The way the intensity of an emotion is apparent from physiological data may be different than the way it is apparent from visual data. Generalizing intensity analysis across different subjects is a challenge yet to

be researched, and it is expected to be a cumbersome problem as different subjects express different levels of emotions in the same situation (Levenson, 1988).

THE BASELINE PROBLEM

When targeting spontaneous behaviour analysis and moving toward real-world settings, one of the basic problems is the Baseline Problem (Nakasone et al., 2005). For tactile modality, The Baseline Problem refers to the problem of finding a condition against which changes in measured physiological signals can be compared—the baseline. For visual modality, The Baseline Problem refers to finding a frame in which the subject is expressionless and against which changes in subject's motion, pose, and appearance can be compared. This is usually achieved by manually segmenting the recordings, or by constraining the recordings to emotional prototypes, or by having the first frame containing baseline/neutral expression. For the audio modality this is usually achieved by segmenting the recordings into turns using energy based Voice Activity Detection and processing each turn separately (e.g., Wollmer et al., 2008). Yet, as pointed out by Levenson (1988) emotion "is rarely superimposed upon a prior state of "rest"; instead, emotion occurs most typically when the organism is in some prior activation." Hence, enforcing existence of expressionless state in each recording or manually segmenting recordings so that each segment contains a baseline expression are strong, unrealistic constrains. This remains a great challenge in automatic analysis, which typically relies on existence of a baseline for analysis and processing of affective information.

DIMENSIONALITY

The space based on which emotions are typically recognized is usually a feature space with a very high dimensionality. For example, Valstar and Pantic (2007) extract 2,520 features for each frame of the input facial video, Wollmer et al. (2008) extract 4,843 features for each utterance, Chanel, Ansari, and Pun (2007) use 16,704 EEG features, Kim (2007) uses 61 features extracted from speech segments and 77 features extracted from bio-signals. The problematic issue here is having fewer training samples than features per sample for learning the target classification, which may lead to under sampling or a singularity problem. To alleviate this problem, dimensionality reduction or feature selection techniques are applied. Linear combination techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and non-linear techniques such as kernel PCA (KPCA) have been used for that purpose (e.g., Chanel et al., 2007; Gunes & Piccardi, 2009; Khan et al., 2009), and so have been feature selection techniques such as Sequential Backward Selection (Kim, 2007). However, how to optimally reduce the dimensionality of continuous multicue and multimodel affect still needs to be explored.

GENERALIZATION

Should automatic affect analysers be able to generalize across subjects or should the recognition be personalized? When it comes to affect recognition from bio-potential signals, the overall amplitudes of the patterns recorded are found to be dependent on the user, suggesting that personalization is required to ensure consistent recognition of significant patterns for these signals (Conati, Chabbal, & Maclaren, 2003). Kim (2007) also found that at times subjects are inconsistent in their emotional expression. Kulic and Croft (2007) reported on the problem of saliency: subjects seem to vary not only in terms of response amplitude and duration, but for some modalities, a number of subjects show no response at all (e.g., only a subset of subjects exhibit heart-rate response). This makes generalization over unseen subjects a very difficult problem. Chanel et al. (2005) emphasize the need of training and evaluating classifiers for each participant separately due

to the aforementioned inter-subject variation. When it comes to other modalities, most of the works in the field report only on subject dependent dimensional affect recognition due to limited number of subjects and data (e.g., Wollmer et al., 2008).

FUSION

For affect sensing and recognition, modality fusion refers to combining and integrating all incoming monomodal events into a single representation of the affect expressed by the user. When it comes to integrating the multiple modalities the major issues are: (i) when to integrate the modalities (i.e., at what abstraction level to do the fusion), and (ii) how to integrate the modalities (i.e., which criteria to use). Typically, the multimodal data fusion is either done at the feature level in a maximum likelihood estimation manner or at the decision level when most of the joint statistical properties (maximum a posteriori) may have been lost (Corradini et al., 2003). To make the multimodal data fusion problem tractable, the individual modalities are usually assumed independent of each other. This simplification allows employing simple parametric models for the joint distributions that cannot capture the complex relationships between the modalities. More importantly, this does not support mutual estimation (e.g., using the audio information to inform the visual information processing; Corradini, Mehta, Bernsen, & Martin, 2003).

The assumption of mutual independence of different modalities is typical for decision-level data fusion. In this approach, a separate classifier processes each modality and the outputs of these classifiers are combined at a later stage to produce the final hypothesis about the shown affective behavior. The decision-level data fusion is the most commonly applied approach in the field, especially when modalities differ in temporal characteristics (e.g., audio and visual modality). Designing optimal strategies for decision-level fusion has been of interest to researchers in the fields of pattern recognition and machine learning, and more recently to researchers in the fields of data mining and knowledge discovery. One approach, which has become popular across many disciplines, is based upon the combination of multiple classifiers, also referred to as an ensemble of experts and/or classifier fusion. For an overview of work done on combining classifiers and for theoretical justification for using simple operators such as majority vote, sum, product, maximum/minimum/median, and adaptation of weights, the readers are referred to the work by Kittler, Hatef, Duin, and Matas (1998). Decision-level data fusion can be obtained at the soft-level (a measure of confidence is associated with the decision), or at the hard-level (the combining mechanism operates on single hypothesis decisions).

Feature-level data fusion is assumed to be appropriate for closely coupled and synchronized modalities (e.g., speech and lip movements). This approach assumes a strict time synchrony between the modalities. Hence, feature-level data fusion tends not to generalize well when the modalities substantially differ in temporal characteristics (e.g., speech and gestures). Therefore, when input from two modalities is fused at the feature level, features extracted from the two modalities should be made synchronous and compatible. The asynchrony between modalities may be of two kinds: (a) asynchrony in subject's signal production (e.g., the facial action might start earlier than the vocalization), and (b) asynchrony in the recording (e.g., video is recorded at 25 Hz, the audio is recorded at 48 kHz, while EEG is recorded at 256-512 Hz). Feature-level fusion becomes more challenging as the number of features increases and when they are of very different natures (e.g., in terms of their temporal properties). Synchronization then becomes of utmost importance. Recent works have attempted synchronization between multiple multimodal cues to support feature-level fusion for the purposes of affect recognition, and reported greater overall accuracy when compared to decision-level fusion (e.g., Gunes & Piccardi, 2009; Shan, Gong, & McOwan,

2007). Gunes and Piccardi (2009) identify the neutral-onset-apex-offset-neutral phases of facial and bodily displays and synchronize the input video sequences at the phase level (i.e., apex phase). Although this method has been used for categorical emotion recognition, if the temporal information and duration of emotions are explicitly modelled, this method can be easily extended to dimensional affect recognition. Savran et al. (2006) have obtained feature/decision level fusion of the fNIRS and EEG feature vectors on a block-by-block basis. In their experiments a block is 12.5 seconds long and represents all emotional stimuli occurring within that time frame. This method can be easily applied to facilitate multimodal dimensional affect recognition. However, choosing an appropriate time-window may pose a challenge.

Outside the affect sensing and recognition field, various techniques have been exploited for implicit data synchronization purposes. For instance, dynamic time warping (DTW) has been used to find the optimal alignment between two time series. This warping between two time series can then be used to find corresponding regions between the two time series and to determine the similarity between them. Variations of Hidden Markov Models (HMM) have also been proposed for this task. Coupled HMM and fused HMM have been used for integrating tightly coupled time series, such as audio and visual features of speech (Pan, Levinson, Huang, & Liang, 2004). Bengio (2004) presented the Asynchronous HMM that could learn the joint probability of pairs of sequences of audiovisual speech data representing the same sequence of events. There are also a number of efforts within the affect sensing and recognition field to exploit the correlation between the modalities and relax the requirement of synchronization by adopting the so-called model-based fusion approach using Bayesian Networks, Multi-stream Fused HMM, tripled HMM, Neural Networks, and so forth. (for details, see Zeng et al., 2009).

Overall, typical reasons to use decision-level fusion (i.e., late integration) instead of feature-level fusion (i.e., early integration) can be summarised as follows (Wu, Oviatt, & Cohen, 1999).

- The feature concatenation used in feature-level fusion results in a high dimensional data space, resulting in a large multimodal dataset.
- Decision-level fusion allows asynchronous processing of the available modalities.
- Decision-level fusion provides greater flexibility in modelling, i.e., it is possible to train different classifiers on different data sources and integrate them without retraining.
- Using decision-level fusion of-the-shelf recognisers can be utilised for single modalities (e.g., speech).
- Decision-level fusion allows adaptive channel weighting between different modalities based on environmental conditions, such as the signal-to-noise ratio.

However, one should note that co-occurrence information (i.e., which multimodal cues co-occur at the same time, which co-occur in time with one occurring after the other, how often are the co-occurrences, etc.) is lost if decision-level fusion is chosen instead of feature-level fusion.

As pointed out by Kim (2007), a user may consciously or unconsciously conceal his or her real emotions as shown by observable cues like facial or vocal expressions, but still reveal them by invisible cues like bio signals. So, how should the fusion proceed when there is conflicting information conveyed by the modalities? This is still an open question that is yet to be investigated. Another issue to consider in affective multimodal data fusion is how to optimally fuse information with high disparity in accuracy (Kim, 2007). In addition, classification methods readily available in machine learning and pattern recognition may not be suitable for emotion-specific problems. The design of emotion-specific classification schemes that can

handle multimodal and spontaneous data is one of the most important issues in the field.

EVALUATION

The evaluation measures applicable to categorical approaches to emotion recognition are not directly applicable to dimensional approaches. For example, Wolmer et al. (2008) use the Mean Squared Error (MSE) between the predicted and the actual value of arousal and valence instead of the recognition rate (i.e., percentage of correctly classified instances). However, whether MSE is the best way to evaluate the performance of dimensional approaches to automatic affect recognition, remains an open issue.

THE STATE-OF-THE-ART

The most commonly employed strategy in automatic dimensional affect classification is to simplify the problem of classifying the six basic emotions to a three-class valence-related classification problem: positive, neutral, and negative emotion classification (e.g., Yu et al., 2008). A similar simplification is to reduce the dimensional emotion classification problem to a two-class problem—positive vs. negative and active vs. passive classification problem—or a four-class problem—quadrants of 2D A-V space classification problem (e.g., Caridakis, Malatesta, Kessous, Amir, Paouzaiou, & Karpouzis, 2006; Fragopanagos & Taylor, 2005). Glowinski et al. (2008), for instance, analyse four emotions, each belonging to one quadrant of the A-V emotion space: high arousal positive valence (joy), high arousal negative valence (anger), low arousal positive valence (relief), and low arousal negative valence (sadness).

Automatic dimensional affect recognition is still in its pioneering stage. It is worth noting that dimensional representation has mostly been used for emotion recognition from physiological signals. Hereby, in Table 2 and Table 3 we briefly summarise automated systems that attempt to model and recognize affect in the continuous dimensional space. This overview is intended to be illustrative rather than exhaustive. Table 2 summarizes representative systems for dimensional affect recognition from a single modality. Table 3 summarizes the utilised classification methods and the performance attained by the methods listed in Table 2. Table 4 summarizes the systems for dimensional affect recognition from multiple modalities. Table 5 summarizes the utilised classification methods and the performance attained by the methods listed in Table 4.

According to the dimensional approach, emotions are represented along a continuum. Therefore, automatic systems adopting this approach should produce continuous values for the target dimensions. Little attention has been paid so far to whether there are definite boundaries along the continuum to distinguish between various levels or intensities. The most common way to explore this issue is to quantize the arousal and valence dimensions into arbitrary number of levels or intensities. Kleinsmith and Bianchi-Berthouze (2007), for instance, use a back-propagation algorithm to build a separate model for each of the affective dimensions for discriminating between levels of affective dimensions from posture (high-low, high-neutral, and low-neutral). Wollmer et al. (2008) use Conditional Random Fields (CRF) for discrete emotion recognition by quantising the continuous labels for valence and arousal to four and/or seven arbitrary levels. Kulic and Croft (2007) perform quantization into 3 categories (low/medium/high), and Chanel et al. (2007) consider 3 classes, namely, excited-negative, excited-positive, and calm-neutral. Karpouzis et al. (2007) focus on positive vs. negative or active vs. passive classes.

The only approach reported in automatic affects sensing field that actually deals with continuous emotions is presented by Wollmer et al. (2008) for emotion recognition from the audio modality. Emotional history is modelled using Long Short-Term Memory Recurrent Networks (LSTM-RNN) which builds upon the principle of recurrent neural networks by including memory cells. LSTM-RNN architecture consists of three layers: an input, a hidden,

Table 2. Overview of the systems for dimensional affect recognition from a single modality

System	Modality/cue	Database	# of Samples	Features	Dimensions
Glowinski et al., 2008	Visual, movement expressivity	Their own	40 portrayals	Gesture dynamics	4 emotions: high arousal (anger and joy) and low arousal (relief and sadness)
Khan et al., 2009	Thermal	Their own (neutral, pretended and evoked facial expressions)	Not reported	facial feature points from images	neutral, positive and negative emotion categories
Kleinsmith and Bianchi-Berthouze, 2007	Visual, static body posture	subjects displaying various body postures given a situation description	111 images	Features from the motion capture system	Valence, arousal, potency, and avoidance
Lee and Narayanan, 2005	Emotional speech	spoken language data obtained from a call centre application	1187 calls, 7200 utterances	a combination of acoustic, lexical, and discourse information	negative and non-negative emotions
Martin, Caridakis, Devillers, Karpouzis, and Abrilian, 2009	Visual, body movement	TV interviews, spontaneous	50 video samples of emotional TV interviews	coarse estimate of the overall movement quantity in a video	emotional activation of a whole video
Shin, 2007	Visual, facial expression images	posed static Korean facial expression database, 6 subjects	287 images	Facial features	pleasure-displeasure and arousal-sleep dimensions
Vogt, André, and Bee, 2008	Emotional speech	Offline speech emotion recognition framework, sentence set in German, 29 students	Not reported	variety of acoustic features like energy, MFCC, pitch and voice quality	positive-active, positive-passive, negative-active, negative-passive mapped on the emotions of joy, satisfaction, anger, frustration
Wollmer et al., 2008	Audio	SAL, 4 subjects	25 recordings, 1,692 turns	variety of acoustic features	positive-active, positive-passive, negative-active, negative-passive

and an output layer, and models long-range dependencies between successive observations. The Long Short-Term Memory cells ensure that events lying back in time are not forgotten. When compared to other classification techniques like Support-Vector Regression, LSTM-RNN achieve a prediction quality which is equal to human performance due to its capability of modelling long range time dependencies.

However, there is no agreement on how to model dimensional emotion space (continuous/quantized) and which classifier is better suited for automatic, multimodal, continuous affect analysis using a dimensional representation. The surveyed works also report a number of

Table 3. The utilised classification methodology and the performance attained by the methods listed in Table 2

System	Classification	Results
Glowinski et al., 2008	Only preliminary analysis no classification reported	Only preliminary analysis no classification reported
Khan et al., 2009	linear discriminants (LDA)	83.3% for posed for 3 classes: neutral, happy and sad; 57.1% for 7 classes; 72% for evoked neutral, happy, sad, disgust and angry.
Kleinsmith and Bianchi-Berthouze, 2007	a back-propagation algorithm with a separate model for each of the 4 affective dimensions	79% for both the valence and arousal, and 81% for both the potency and avoidance dimensions
Lee and Narayanan, 2005	discriminant classifiers (LDC) with Gaussian class-conditional probability and k-nearest neighbourhood classifiers (k-NN) to detect negative versus non-negative emotions	Improvement of 40.7% for males and 36.4% for females via fusion of information
Martin et al., 2009	discriminant analysis	67.2% for pretended, and 72% for evoked expressions of neutral, happy, disgusted, surprised, and angry emotions
Shin, 2007	a 3-layer neural network with 2 output nodes of pleasure-displeasure and arousal-sleep	Only coarse comparison btw. NN and mean A-V human annotation
Vogt et al., 2008	Naive Bayes and support vector machine classifiers to distinguish between the four quadrants of the A-V space	an average of 55% for a 4 class problem
Wollmer et al., 2008	Long Short-Term Memory Recurrent Neural Net, Support Vector Machines, Conditional Random Fields, and Support Vector Regressor	0.18 MSE using speaker dependent validation

additional challenging issues as summarized in Table 6.

Overall, automatic human affect recognition based on a dimensional approach is still in its infancy. As can be seen from Tables 2-5, the comparison of results attained by different surveyed systems is difficult to conduct as systems use different training/testing datasets (which differ in the way emotions are elicited and annotated), they differ in the underlying model of emotions (i.e., target emotional categories) as well as in the employed modality or combination of modalities and the applied evaluation method (Arroyo-Palacios & Romano, 2008). Wagner et al. (2005) argue that for the current multimodal affect recognizers, the achieved recognition rates depend on the type of the utilized data, and whether the emotions were acted or not, rather than on the used algorithms and classification methods. All of this makes it difficult to quantitatively and comparatively evaluate the accuracy of the A-V modelling and the effectiveness of the developed systems.

As a consequence, it remains unclear which classification method is suitable for dimensional affect recognition from which modalities and cues. Opportunities for solving this problem can be potentially searched in other relevant research fields. For example, the A-V dimensional approach has been mostly used for affective content classification from music or videos (e.g., Xu et al., 2008; Zhang et al., 2008). Therefore, methodologies in these fields seem more mature and advanced compared to those in automatic human affect recognition field. Zhang et al. (2008), for instance, perform affec-

Table 4. Overview of the systems for dimensional affect recognition from multiple modalities

System	Modality/cue	Database	# of Samples	Features	Dimensions
Caridakis, Karpouzis, and Kollias, 2008	facial expression, body gestures and audio	SAL, 4 subjects	Not reported	Various visual and acoustic features	neutral and four A-V quadrants
Chanel et al., 2007	Tactile, physiological	Their own, 1 subject, recall of past emotional events	Not reported	EEG and peripheral features	arousal and valence
Forbes-Riley and Litman, 2004	audio and text	student emotions from tutorial spoken dialogues	Not reported	variety of acoustic and prosodic, text-based, and contextual features	negative, neutral and positive emotions
Haag et al., 2004	Tactile, physiological	Their own, 1 subject	1000 samples	heart rate, BVP, EMG, skin conductivity, respiration	arousal and valence
Karpouzis et al., 2007	facial expression, body gestures and/or audio	SAL, 4 subjects	76 Passages, 1600 tunes	Various visual and acoustic features	negative vs. positive, active vs. passive
Kim, 2007	speech and physiological signals	A corpus of spontaneous vocal and physiological emotions, using a modified version of the quiz "Who wants to be a millionaire?", 3 subjects	343 samples	EMG, SC, ECG, BVP, Temp, RSP and acoustic features	either of the four A-V quadrants
Kulic and Croft, 2007	Tactile, physiological	Their own, context of human-robot interaction, 36 subjects	2-3 examples for each affect category	heart rate, perspiration rate, and facial muscle contraction	6 affect categories (low/medium/high-valence/arousal)
Wagner, Kim, and Andre, 2005	Tactile, physiological	Their own, 1 subject listening to songs	25 recordings for each emotion	physiological signals	negative (anger/sadness), positive (joy/pleasure), valence and high arousal (joy/anger), low arousal (sadness/pleasure)

tive video content analysis from MTV clips in terms of A-V space by employing a clustering method called Affinity Propagation (AP). The main reason for this choice is the fact that they do not have apriori knowledge of how many affective categories a classifier should output.

Table 5. The utilised classification methodology and the performance attained by the methods listed in Table 4

System	Classification	Explicit Fusion	Results
Caridakis et al., 2008	a feed-forward back-propagation network to map tunes into either of the 4 A-V quadrants or the neutral state	Not reported	reported as reduced MSE for every tune
Chanel et al., 2007	linear discriminant analysis (LDA) and support vector machine (SVM)	Not reported	67% accuracy for 3 classes (negatively excited, positively excited, and calm-neutral), and 79% accuracy for 2 classes (negatively vs. positively excited) using EEG, 53% accuracy for 3 classes and 73% accuracy for 2 classes using peripheral signals
Forbes-Riley and Litman, 2004	AdaBoost to boost a decision tree algorithm for negative, neutral and positive emotions	Not reported	84.75% for a 3 class problem
Haag et al., 2004	separate network for valence and arousal, each with a single output node corresponding to the valence or arousal value	Not reported	96.6% for arousal, 89.9% for valence.
Karpouzis et al., 2007	a Simple Recurrent Network that outputs either of the 4 classes (3 for the possible emotion quadrants, one for neutral affective state)	Not described	67% recognition accuracy using the visual modality and 73% using prosody, 82% after fusion (whether on unseen subject/data is not specified)
Kim, 2007	modality-specific LDA-based classification; a hybrid fusion scheme where the output of feature-level fusion is fed as an auxiliary input to the decision-level fusion stage	Decision level fusion and hybrid fusion by integrating results from feature and decision level fusion	51% for bio-signals, 54% for speech, 55% applying feature fusion, 52% for decision fusion, and 54% for hybrid fusion, subject independent validation.
Kulic and Croft, 2007	3 HMMs for valence (low, medium, and high) and 3 HMMs for arousal (low, medium, and high)	Not reported	an accuracy of 64% for novel data
Lee and Narayanan, 2005	discriminant classifiers (LDC) with Gaussian class-conditional probability and k-nearest neighbourhood classifiers (k-NN) to detect negative versus non-negative emotions	Decision level fusion	Improvement of 40.7% for males and 36.4% for females via fusion of information
Wagner et al., 2005	k-nearest neighbour (kNN), linear discriminant function (LDF) and a multilayer perceptron (MLP) to recognize 4 emotion classes	Not reported	High vs. low arousal 95%, and negative vs. positive 87%

Table 6. Reported challenges for dimensional affect recognition

System	Challenges Encountered
Chanel et al., 2007	EEG signals are good for valence assessment. Peripheral signals better correlate with arousal than with valence. Peripheral signals appear to be appropriate for modelling calm-neutral vs. excited dimension, but are problematic for the negative vs. positive dimension.
Haag et al., 2004	Estimation of valence is harder than estimation of arousal.
Karpouzis et al., 2007	Disagreement (frame-based) between human observers (annotators) affects the performance of the automated systems. The system should take into account the inter-observer disagreement, by comparing this to the level of disagreement between the ground truth and the results attained by the system.
Kim, 2007	Recognition is subject and modality dependant.
Kulic and Croft, 2007	There is a considerable inter-subject variability in the signal amplitude and its length. Hence, it is hard to develop a system that can perform well for all subjects and generalize well for unseen subjects.

Another good example on how to handle data comprising continuous values comes again from the affective content analysis field. Yang et al. (2007) model emotions as continuous variables composed of arousal and valence values, and formulate music emotion recognition as a regression problem. This choice is based on the fact that the regression approach is inherently continuous, and exhibits promising prediction accuracy; it learns the predicting rules according to the ground truth and, if categorical description is needed, the regression results can be easily converted to binary or quaternary results. Various types of regressors can be used for this task: the multiple linear regression (MLR), support vector regression (SVR), and AdaBoost.RT, etc. The ground truth is obtained by averaging subjects' opinions about the A-V values for each input sample. The emotion plane is viewed as a coordinate space spanned by the A-V values (each value confined within [-1, 1]). Then Yang et al., train two regressors to predict the A-V values. The arousal and valence models are weighted combinations of some component functions, which are computed along the timeline. Yang et al. (2007) train the two regressors separately under the assumption that the correlation between arousal and valence is embedded in the ground truth. Although the context is different from that of human affect sensing, affect recognition researchers could potentially benefit from the aforementioned methodologies.

There exist a number of studies that focus on dimensional modelling of affect in the context of empathic companions (e.g., Nakasone et al., 2005), educational games (e.g., Conati et al., 2003), game interfaces (Kim et al., 2004), and speech analysis (Jin & Wang, 2005). Although interesting as the first attempts toward application-oriented systems, these works are usually based on manual analysis and do not attempt automatic dimensional affect recognition.

In summary, the issues pertinent in dimensional affect recognition include reliability of the ground-truth, determining duration of emotions for automatic analysis, determining the baseline, dimensionality reduction, modelling intensity of emotions, high inter-subject variation, defining optimal fusion of cues/modalities, and identifying appropriate classification methods and evaluation measures.

CONCLUSION AND DISCUSSION

This article discussed the problem domain of affect sensing using a dimensional approach and explored the current state-of-the-art in continuous, dimensional affect recognition.

The analysis provided in this article indicates that the automatic affect analysis field has slowly started shifting from categorical emotion recognition to dimensional emotion recognition. Existing dimensional affect analysis systems mostly deal with spontaneous data obtained in less-controlled environments (i.e., subjects are taking part in interactions, subjects are not always stationary, etc.), and can handle a small number of (quantized) affective dimension categories. However, note that real-world settings pose many challenges to affect sensing and recognition (Conati et al., 2003). Firstly, it is not easy to obtain a high level of reliability among independent observers annotating the affect data. In addition, when subjects are not restricted in terms of mobility, the level of noise in all recorded signals tends to increase. This is particularly the case for bio-signals. No solution has yet been proposed to solve these problems.

In general, modelling emotions continuously using the dimensions of arousal and valence is not a trivial problem as these dimensions are not universally perceived and understood by human observers. It seems that the perception of arousal is more universal than is the perception of valence (Zhang et al., 2008). Similar findings have been reported by Kleinsmith and Bianchi-Berthouze (2007), who found that ratings of arousal contained very small variability among different observers, when body postures were mapped onto affective dimensions. Also, for audio modality variability of ratings of arousal appears to be smaller than that of valence (Jin & Wang, 2005). Wolmer et al. (2008) also reported that automatic analysis results for activation/arousal are remarkably better than those for valence when using audio information. Yet, valence appears to be more stable than arousal in dimensional facial expression recognition from static images (Shin, 2007). Having said the above, it can be concluded that stability of inter-observer agreement on valence and arousal is highly dependent on the modality employed. Hence, this makes the problem of obtaining a reliable ground truth for multimodal recordings a true challenge.

To address this problem Kim (2007) suggests that emotion recognition problem should be decomposed into several processes. One stage could be recognizing arousal through physiological channels, while recognizing valence via audiovisual channels. The second stage can then be resolving uncertainties between adjacent emotion classes in the 2D space by cumulative analysis of user's context information. A more thorough investigation is needed to test this suggestion and propose a similar set of processes to be applied when other cues and modalities are employed.

One of the main disadvantages of biopotential-based affect recognition systems is the fact that they are cumbersome and invasive and require placing sensors physically on the human body (e.g., a sensor clip that is mounted on subject's earlobe, a BCI mounted on the subject's head, etc.; Takahashi, 2004). Moreover, EEG has been found to be very sensitive to electrical signals emanating from facial muscles while emotions are being expressed via face. Therefore, in a multimodal affect recognition system, simultaneous use of these modalities needs to be reconsidered. Additionally, during recordings, the fNIRS device is known to cover the eyebrows. This in turn poses another challenge: facial features occlusion. However, new forms of non-contact physiological sensing might facilitate better utilisation of psychological signals as input to multimodal affect recognition systems.

To the best of our knowledge, to date, only a few systems have been reported that actually achieved dimensional affect recognition from multiple modalities. These are summarised in Tables 4 and 5. Further efforts are needed to identify the importance and feasibility of the following important issues.

- Among the available remotely observable and remotely unobservable modalities, which ones should be used for automatic dimensional affect recognition? Does this depend on the context? Will the recognition accuracy increase as the number of

modalities a system can analyse increases?
- Kim (2007) found that speech and physiological data contain little complementary information. Accordingly, should we use equal weights for each modality or should we investigate the innate priority among the modalities to be preferred for each emotional dimension/state?
- Chanel et al. (2005) report that although it is difficult to self-assess arousal, using classes generated from self-assessment of emotions facilitate greater accuracy in recognition. When labelling emotions, should one use self assessment or independent observer's assessment? Which signals better correlate with self assessment and which ones correlate with independent observer assessment?
- How does *the baseline problem* affect recognition? Is an objective *basis* (e.g., a frame containing an expressionless display) strictly needed prior to computing the arousal and valence values? If so, how can this be obtained in a fully automatic manner from spontaneous data?
- Considering the fact that different emotions may have similar or identical valence or arousal values (Haag et al., 2004), should the affect recognizers attempt to recognize distinct emotion categories rather than A-V intensities? Does this depend on the context? How should affective states be mapped onto the A-V space? Should we follow a hierarchical framework where similar affective states are grouped into the same category?
- How should intensity be modelled for dimensional and continuous affect recognition? Should the aim be personalizing systems for each subject, or creating systems that are expected to generalize across subjects?
- In a continuous emotional space, how should duration of emotion be defined? How can this be incorporated in automated systems? Will focusing on shorter or longer observations affect the accuracy of the recognition process?
- In real-world uncontrolled settings it is very difficult to elicit balanced amount of data for each emotion dimension to be elicited. For instance, a bias toward quadrant 1 (positive arousal, positive valence) exists in the SAL database portion used by (Caridakis et al., 2008). So, how should the issue of unbalanced data/classes inherent to real-world settings (Chanel et al., 2005) be handled?

The most notable issue in the field is the existence of a gap between different communities. Machine affect recognition community seems to use different databases compared to psychology and cognitive sciences communities. Also for annotation of the data, a more uniform and multi-purpose scheme that can accommodate all possible research aims, modalities and cues should be explored.

The systems surveyed in this article represent initial but crucial steps toward finding solutions to the aforementioned problems, and realization of automatic, multimodal, dimensional and continuous recognition of human affect.

ACKNOWLEDGMENT

Current research of Hatice Gunes is funded by the European Community's 7th Framework Programme [FP7/2007-2013] under the grant agreement no 211486 (SEMAINE). The work of Maja Pantic is funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007-StG-203143 (MAHNOB).

REFERENCES

Aftanas, L. I., Pavlov, S. V., Reva, N. V., & Varlamov, A. A. (2003). Trait anxiety impact on the EEG theta band power changes during appraisal of threatening and pleasant visual stimuli. *International Journal of Psychophysiology*, *50*(3), 205–212. doi:10.1016/S0167-8760(03)00156-9

Aftanas, L. I., Reva, N. V., Varlamov, A. A., Pavlov, S. V., & Makhnev, V. P. (2004). Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: Temporal and topographic characteristics. *Neuroscience and Behavioral Physiology*, *34*(8), 859–867. doi:10.1023/B:NEAB.0000038139.39812.eb

Ambadar, Z., Schooler, J., & Cohn, J. F. (2005). Deciphering the enigmatic face: The importance of facial dynamics in interpreting subtle facial expressions. *Psychological Science*, *16*(5), 403–410. doi:10.1111/j.0956-7976.2005.01548.x

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behaviour as predictors of interpersonal consequences: A meta–analysis. *Psychological Bulletin*, *11*(2), 256–274. doi:10.1037/0033-2909.111.2.256

Argyle, M. (1975). *Bodily communication*. London: Methuen.

Arroyo-Palacios, J., & Romano, D. M. (2008, August). Towards a standardization in the use of physiological signals for affective recognition systems. In *Proceedings of Measuring Behavior 2008,* Maastricht, The Netherlands (pp. 121-124). Noldus.

Banziger, T., & Scherer, K. R. (2007, September) Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus. In A. Paiva, R. Prada, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction,* Lisbon, Portugal (LNCS 4738, pp. 476-487).

Baron-Cohen, S., & Tead, T. H. E. (2003) *Mind reading: The interactive guide to emotion*. London: Jessica Kingsley Publishers.

Batliner, A., Fischer, K., Hubera, R., Spilkera, J., & Noth, E. (2003). How to find trouble in communication. *Speech Communication*, *40*, 117–143. doi:10.1016/S0167-6393(02)00079-1

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40*(2), 531–539. doi:10.3758/BRM.40.2.531

Bengio, S. (2004). Multimodal speech processing using asynchronous hidden markov models. *Information Fusion*, *5*, 81–89. doi:10.1016/j.inffus.2003.04.001

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*, 119–155. doi:10.1016/S1071-5819(03)00018-1

Buller, D., Burgoon, J., White, C., & Ebesu, A. (1994). Interpersonal deception: Vii. Behavioural profiles of falsification, equivocation and concealment. *Journal of Language and Social Psychology*, *13*(5), 366–395. doi:10.1177/0261927X94134002

Campbell, N., & Mokhtari, P. (2003, August). Voice quality: The 4th prosodic dimension. In *Proceedings of the International Congress of Phonetic Sciences,* Barcelona (pp. 2417-2420).

Camras, L. A., Meng, Z., Ujiie, T., Dharamsi, K., Miyake, S., & Oster, H. (2002). Observing emotion in infants: Facial expression, body behaviour, and rater judgments of responses to an expectancy-violating event. *Emotion (Washington, D.C.)*, *2*, 179–193. doi:10.1037/1528-3542.2.2.179

Camurri, A., Mazzarino, B., & Volpe, G. (2003, April) Analysis of expressive gesture: The EyesWeb expressive gesture processing library. In *Proceedings of the Gesture Workshop,* Genova, Italy (pp. 460-467).

Caridakis, G., Karpouzis, K., & Kollias, S. (2008). User and context adaptive neural networks for emotion recognition. *Neurocomputing*, *71*, 13–15, 2553–2562. doi:10.1016/j.neucom.2007.11.043

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaiou, A., & Karpouzis, K. (2006, November). Modelling naturalistic affective states via facial and vocal expression recognition. In *Proceedings 8th ACM International Conference on Multimodal Interfaces (ICMI '06),* Banff, Alberta, Canada (pp. 146-154). ACM Publishing.

Chanel, G., Ansari-Asl, K., & Pun, T. (2007, October). Valence-arousal evaluation using physiological signals in an emotion recall paradigm. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics,* Montreal, Quebec, Canada (pp. 2662-2667). Washington, DC: IEEE Computer Society.

Chanel, G., Kronegg, J., Grandjean, D., & Pun, T. (2002). *Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals* (Tech. Rep. 05.02). Geneva, Switzerland: Computer Vision Group, Computing Science Center, University of Geneva.

Changchun, L., Rani, P., & Sarkar, N. (2005, August). An empirical study of machine learning techniques for affect recognition in human-robot interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems,* Edmonton, Canada (pp. 2662-2667). Washington, DC: IEEE Computer Society.

Conati, C., Chabbal, R., & Maclaren, H. A. (2003, June). *Study on using biometric sensors for monitoring user emotions in educational games*. Paper presented at the Workshop on Assessing and Adapting to User Attitudes and Affect: Why, When and How? User Modelling (UM-03), Johnstown, PA.

Corradini, A., Mehta, M., Bernsen, N. O., & Martin, J.-C. (2003, August). *Multimodal input fusion in human computer interaction on the example of the ongoing nice project*. In *Proceedings of the NATO: Asi Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management,* Tsakhkadzor, Armenia (pp. 223-234).

Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Nonverbal Behavior*, *28*(2), 117–139. doi:10.1023/B:JONB.0000023655.25550.be

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schroder, M. (2000, September). 'FEELTRACE': An instrument for recording perceived emotion in real time. In *Proceedings of the ISCA Workshop on Speech and Emotion,* Belfast, Northern Ireland (pp. 19-24).

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., & Fellenz, W. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80. doi:10.1109/79.911197

Darwin, C. (1998). *The expression of the emotions in man and animals* (3rd ed.). New York: Oxford University Press.

Davitz, J. (1964). Auditory correlates of vocal expression of emotional feeling. In J. Davitz (Ed.), The communication of emotional meaning (pp. 101-112). New York: McGraw-Hill.

De Silva, P. R. S., Osano, M., Marasinghe, A., & Madurapperuma, A. P. (2006, April). Towards recognizing emotion with affective dimensions through body gestures. In *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition,* Southampton, UK (pp. 269-274).

DePaulo, B. (2003). Cues to deception. *Psychological Bulletin*, *129*(1), 74–118. doi:10.1037/0033-2909.129.1.74

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., et al. (2007, September). The HUMAINE Database: addressing the needs of the affective computing community. In *Affective Computing and Intelligent Interaction: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction,* Lisbon, Portugal (LNCS 4738, pp. 488-500).

Dreuw, P., Deselaers, T., Rybach, D., Keysers, D., & Ney, H. (2006, April). Tracking using dynamic programming for appearance-based sign language recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition,* Southampton, UK (pp. 293-298). Washington, DC: IEEE Computer Society.

Driver, J., & Spence, C. (2000). Multisensory perception: Beyond modularity and convergence. *Current Biology*, *10*(20), 731–735. doi:10.1016/S0960-9822(00)00740-5

Ekman, P. (1982). *Emotion in the human face*. Cambridge, UK: Cambridge University Press.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences*, *1000*, 105–221.

Ekman, P., & Friesen, W. V. (1967). Head and body cues in the judgment of emotion: A reformulation. *Perceptual and Motor Skills*, *24*, 711–724.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face: A guide to recognizing emotions from facial clues*. Englewood Cliffs, NJ: Prentice-Hall.

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., & Hager, J. C. (2002). *Facial action coding system*. Salt Lake City, UT: A Human Face.

El Kaliouby, R., & Robinson, P. (2005, June 27-July 2). Real-time inference of complex mental states from facial expressions and head gestures. In *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW 2004),* Washington, DC (Vol. 10, pp. 154). Washington, DC: IEEE Computer Society.El Kaliouby, R., & Teeters, A. (2007, November). Eliciting, capturing and tagging spontaneous facial affect in autism spectrum disorder. In *Proceedings of the 9th International Conference on Multimodal Interfaces,* Nagoya, Japan (pp. 46-53).

Elgammal, A., Shet, V., Yacoob, Y., & Davis, L. S. (2003, June). Learning dynamics for exemplar-based gesture recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Madison, WI (pp. 571-578). Washington, DC: IEEE Computer Society.

Fasel, I. R., Fortenberry, B., & Movellan, J. R. (2005). A generative framework for real-time object detection, and classification. *Computer Vision and Image Understanding*, *98*(1), 182–210. doi:10.1016/j.cviu.2004.07.014

Forbes-Riley, K., & Litman, D. (2004, May). Predicting emotion in spoken dialogue from multiple knowledge sources. In *Proceedings of the Human Language Technology Conference North America Chapter of the Association for Computational Linguistics (HLT-NAACL 2004),* Boston (pp. 201-208).

Fragopanagos, F., & Taylor, J. G. (2005). Emotion recognition in human-computer interaction. *Neural Networks*, *18*, 389–405. doi:10.1016/j.neunet.2005.03.006

Friesen, W. V., & Ekman, P. (1984). *EMFACS-7: Emotional facial action coding system* (unpublished manual). San Francisco: University of California, San Francisco.

Glowinski, D., Camurri, A., Volpe, G., Dael, N., & Scherer, K. (2008, June). Technique for automatic emotion recognition by body gesture analysis. In *Proceedings of the 2008 Computer Vision and Pattern Recognition Workshops,* Anchorage, AK (pp. 1-6). Washington, DC: IEEE Computer Society.

Grandjean, D., Sander, D., & Scherer, K. R. (2008). Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and Cognition*, *17*(2), 484–495. doi:10.1016/j.concog.2008.03.019

Grimm, M., Kroschel, K., & Narayanan, S. (2008, June). The Vera am Mittag German audio-visual emotional speech database. In *Proceedings of the IEEE International Conference on Multimedia and Expo,* Hannover, Germany (pp. 865-868). Washington, DC: IEEE Computer Society.

Gross, M. M., Gerstner, G. E., Koditschek, D. E., Fredrickson, B. L., & Crane, E. A. (2006). *Emotion recognition from body movement kinematics*. Retrieved from http://sitemaker.umich.edu/mgrosslab/files/abstract.pdf

Gunes, H., & Piccardi, M. (2006, October). Creating and annotating affect databases from face and body display: A contemporary survey. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics,* Taipei, Taiwan (pp. 2426-2433).

Gunes, H., & Piccardi, M. (2008). From mono-modal to multi-modal: Affect recognition using visual modalities. In D. Monekosso, P. Remagnino, & Y. Kuno (Eds.), *Ambient intelligence techniques and applications* (pp. 154-182). Berlin, Germany: Springer-Verlag.

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, 39*(1), 64-84.

Gunes, H., Piccardi, M., & Pantic, M. (2008). From the lab to the real world: Affect recognition using multiple cues and modalities. In Jimmy Or (Ed.), *Affective computing, focus on emotion expression, synthesis and recognition* (pp. 185-218). Vienna, Austria: I-Tech Education and Publishing.

Haag, A., Goronzy, S., Schaich, P., & Williams, J. (2004, June). Emotion recognition using bio-sensors: First steps towards an automatic system. In E. André, L. Dybkjær, W. Minker, & P. Heisterkamp (Eds.), *Affective Dialogue Systems: Tutorial and Research Workshop (ADS 2004),* Kloster Irsee, Germany (LNCS 3068, pp. 36-48).

Hadjikhani, N., & De Gelder, B. (2003). Seeing fearful body expressions activates the fusiform cortex and amygdala. *Current Biology*, *13*, 2201–2205. doi:10.1016/j.cub.2003.11.049

Jin, X., & Wang, Z. (2005, October). An emotion space model for recognition of emotions in spoken chinese. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII 2005),* Beijing, China, (pp. 397-402).

Juslin, P. N., & Scherer, K. R. (2005). Vocal expression of affect. In J. Harrigan, R. Rosenthal, & K. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research* (pp. 65-135). Oxford, UK: Oxford University Press.

Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., et al. (2007, November). Modelling naturalistic affective states via facial, vocal and bodily expressions recognition. In J. G. Carbonell & J. Siekmann (Eds.), *Artifical Intelligence for Human Computing: ICMI 2006 and IJCAI 2007 International Workshops,* Banff, Canada (LNAI 4451, pp. 92-116).

Keltner, D., & Ekman, P. (2000). Facial expression of emotion. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (pp. 236-249). New York: Guilford Press.

Khan, M. M., Ingleby, M., & Ward, R. D. (2006). Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations. *ACM Transactions on Autonomous and Adaptive Systems*, *1*(1), 91–113. doi:10.1145/1152934.1152939

Khan, M. M., Ward, R. D., & Ingleby, M. (2006, June). Infrared thermal sensing of positive and negative affective states. In *Proceedings of the IEEE Conference on Robotics, Automation and Mechatronics,* Bangkok, Thailand (pp. 1-6). Washington, DC: IEEE Computer Society.

Khan, M. M., Ward, R. D., & Ingleby, M. (2009). Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. *ACM Transactions on Applied Perception*, *6*(1), 6. doi:10.1145/1462055.1462061

Kim, J. (2007). Bimodal emotion recognition using speech and physiological changes. In M. Grimm, K. Kroschel (Eds.), *Robust speech recognition and understanding* (pp. 265-280). Vienna, Austria: I-Tech Education and Publishing.

Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(3), 226–239. doi:10.1109/34.667881

Kleinsmith, A., & Bianchi-Berthouze, N. (2007, September). Recognizing affective dimensions from body posture. In *Affective Computing and Intelligent Interaction: 2nd International Conference,* Lisbon, Portugal (LNCS 4738, pp. 48-58).

Kleinsmith, A., Ravindra De Silva, P., & Bianchi-Berthouze, N. (2005, October) Grounding affective dimensions into posture features. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII 2005),* Beijing, China (pp. 263-270).

Kleinsmith, A., Ravindra De Silva, P., & Bianchi-Berthouze, N. (2006). Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers*, *18*, 1371–1389. doi:10.1016/j.intcom.2006.04.003

Kulic, D., & Croft, E. A. (2007). Affective state estimation for human–robot interaction. *IEEE Transactions on Robotics*, *23*(5), 991–1000. doi:10.1109/TRO.2007.904899

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, *13*(2), 293–303. doi:10.1109/TSA.2004.838534

Levenson, R. W. (1988). Emotion and the autonomic nervous system: A prospectus for research on autonomic specificity. In H. L. Wagner (Ed.), *Social psychophysiology and emotion: Theory and clinical applications* (pp. 17-42). New York: John Wiley & Sons

Lienhart, R., & Maydt, J. (2002, September). An extended set of hair-like features for rapid object detection. In *Proceedings of the IEEE International Conference on Image Processing,* New York (Vol. 1, pp. 900-903). Washington, DC: IEEE Computer Society.

Littlewort, G. C., Bartlett, M. S., & Lee, K. (2007, November). Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the 9th International Conference on Multimodal Interfaces,* Nagoya, Japan (pp. 15-21). ACM Publishing.

Martin, J.-C., Caridakis, G., Devillers, L., Karpouzis, K., & Abrilian, S. (2009). Manual annotation and automatic image processing of multimodal emotional behaviours: Validating the annotation of TV interviews. *Personal and Ubiquitous Computing*, *13*(1), 69–76. doi:10.1007/s00779-007-0167-y

Mehrabian, A., & Russell, J. (1974). *An approach to environmental psychology*. Cambridge, MA: MIT Press.

Nakasone, A., Prendinger, H., & Ishizuka, M. (2005, September). Emotion recognition from electromyography and skin conductance. In Proceedings of the 5th International Workshop on Biosignal Interpretation, Tokyo (pp. 219-222).

Nakayama, K., Goto, S., Kuraoka, K., & Nakamura, K. (2005). Decrease in nasal temperature of rhesus monkeys (Macaca mulatta) in negative emotional state. *Journal of Physiology and Behavior*, *84*, 783–790. doi:10.1016/j.physbeh.2005.03.009

Ning, H., Han, T. X., Hu, Y., Zhang, Z., Fu, Y., & Huang, T. S. (2006, April). A real-time shrug detector. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition,* Southampton, UK (pp. 505-510). Washington, DC: IEEE Computer Society.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, *97*, 315–331. doi:10.1037/0033-295X.97.3.315

Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning.* Chicago: University of Illinois Press.

Pan, H., Levinson, S. E., Huang, T. S., & Liang, Z.-P. (2004). A fused hidden markov model with application to bimodal speech processing. *IEEE Transactions on Signal Processing*, *52*(3), 573–581. doi:10.1109/TSP.2003.822353

Pantic, M., & Bartlett, M. S. (2007). Machine analysis of facial expressions. In K. Delac & M. Grgic (Eds.), *Face recognition* (pp. 377-416). Vienna, Austria: I-Tech Education and Publishing.

Pantic, M., Nijholt, A., Pentland, A., & Huang, T. (2008). Human-centred intelligent human-computer interaction (HCI2): How far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, *1*(2), 168–187. doi:10.1504/IJAACS.2008.019799

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. (2007). Machine understanding of human behaviour. In *Artifical Intelligence for Human Computing* (LNAI 4451, pp. 47-71).

Pantic, M., & Rothkrantz, L. J. M. (2003). Towards an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, *91*(9), 1370–1390. doi:10.1109/JPROC.2003.817122

Pavlidis, I. T., Levine, J., & Baukol, P. (2001, October). Thermal image analysis for anxiety detection. In *Proceedings of the International Conference on Image Processing,* Thessaloniki, Greece (Vol. 2, pp. 315-318). Washington, DC: IEEE Computer Society.

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. doi:10.1109/34.954607

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. In K. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 197-219). Hillsdale, NJ: Lawrence Erlbaum Associates.

Poppe, R. (2007). Vision-based human motion analysis: An overview. *Computer Vision and Image Understanding*, *108*(1-2), 4–18. doi:10.1016/j.cviu.2006.10.016

Pun, T., Alecu, T. I., Chanel, G., Kronegg, J., & Voloshynovskiy, S. (2006). Brain-computer interaction research at the computer vision and multimedia laboratory, University of Geneva. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *14*(2), 210–213. doi:10.1109/TNSRE.2006.875544

Puri, C., Olson, L., Pavlidis, I., Levine, J., & Starren, J. (2005, April). StressCam: Non-contact measurement of users' emotional states through thermal imaging. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 2005),* Portland, OR (pp. 1725-1728). ACM Publishing.

Riseberg, J., Klein, J., Fernandez, R., & Picard, R. W. (1998, April). Frustrating the user on purpose: Using biosignals in a pilot study to detect the user's emotional state. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI 1998),* Los Angeles (pp. 227-228). ACM Publishing.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178. doi:10.1037/h0077714

Russell, J. A. (1997). Reading emotions from and into faces: resurrecting a dimensional contextual perspective. In J. A. Russell & J. M. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 295-320). New York: Cambridge University Press.

Russell, J. A., & Fernández-Dols, J. M. (Eds.). (1997). *The psychology of facial expression.* New York: Cambridge University Press.

Savran, A., Ciftci, K., Chanel, G., Mota, J. C., Viet, L. H., Sankur, B., et al. (2006, July 17-August 11). Emotion detection in the loop from brain signals and facial images. In *Proceedings of eNTERFACE 2006,* Dubrovnik, Croatia. Retrieved from http://www.enterface.net

Scherer, K. R. (2000). Psychological models of emotion. In J. Borod (Ed.), *The neuropsychology of emotion* (pp. 137-162). New York: Oxford University Press.

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. New York: Oxford University Press.

Schmidt, K. L., & Cohn, J. F. (2001). Human facial expressions as adaptations: Evolutionary questions in facial expression research. *Yearbook of Physical Anthropology*, *44*, 3–24. doi:10.1002/ajpa.20001

Shan, C., Gong, S., & McOwan, P. W. (2007, September). *Beyond facial expressions: Learning human emotion from body gestures*. Paper presented at the British Machine Vision Conference, Warwick, UK.

Shin, Y. (2007, May). Facial expression recognition based on emotion dimensions on manifold learning. In *Proceedings of International Conference on Computational Science,* Beijing, China (Vol. 2, pp. 81-88).

Takahashi, K. (2004, December). Remarks on emotion recognition from multi-modal bio-potential signals. In *Proceedings of the IEEE International Conference on Industrial Technology,* Hammamet, Tunisia (pp. 1138-1143).

Tian, Y. L., Kanade, T., & Cohn, J. F. (2002, May). Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Proceedings of the IEEE International Conference on Automaitc Face and Gesture Recognition,* Washington, DC (pp. 218-223). Washington, DC: IEEE Computer Society.

Tomkins, S. S. (1962). *Affect, imagery, consciousness: Vol. 1. The positive affects*. New York: Springer.

Tomkins, S. S. (1963). *Affect, imagery, consciousness. Vol. 2: The negative affects*. New York: Springer.

Tsiamyrtzis, P., Dowdall, J., Shastri, D., Pavlidis, I., Frank, M. G., & Ekman, P. (2007). Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, *71*(2), 197–214. doi:10.1007/s11263-006-6106-y

Valstar, M. F., Gunes, H., & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the 9th International Conference on Multimodal Interfaces,* Nagoya, Japan (pp. 38-45). ACM Publishing.

Valstar, M. F., & Pantic, M. (2007, October). Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In M. Lew, N. Sebe, T. S. Huang, E. M. Bakker (Eds.), *Human–Computer Interaction: IEEE International Workshop, HCI 2007,* Rio de Janeiro, Brazil (LNCS 4796, pp. 118-127).

Van den Stock, J., Righart, R., & De Gelder, B. (2007). Body expressions influence recognition of emotions in the face and voice. *Emotion (Washington, D.C.)*, *7*(3), 487–494. doi:10.1037/1528-3542.7.3.487

Vianna, D. M., & Carrive, P. (2005). Changes in cutaneous and body temperature during and after conditioned fear to context in the rat. *The European Journal of Neuroscience*, *21*(9), 2505–25012. doi:10.1111/j.1460-9568.2005.04073.x

Villalba, S. D., Castellano, G., & Camurri, A. (2007, September). Recognising human emotions from body movement and gesture dynamics. In *Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction (ACII 2007),* Lisbon, Portugal (pp. 71-82).

Vinciarelli, A., Pantic, M., & Bourlard, H. (2009, December). Social signal processing: Survey of an emerging domain. *Image and Vision Computing Journal,* 27(12), 1743-1759.Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Kauai, HI (Vol. 1, pp. 511-518).

Vogt, T., André, E., & Bee, N. (2008, June). EmoVoice—a framework for online recognition of emotions from voice. In *Perception in Multimodal Dialogue Systems: 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems (PIT 2008),* Kloster Irsee, Germany (LNCS 5078, pp. 188-199).

Wagner, J., Kim, J., & Andre, E. (2005, July). From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo,* Amsterdam, The Netherlands (pp. 940-943). Washington, DC: IEEE Computer Society.

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviours: Differentiation of multimodal information. *Psychological Bulletin*, *121*(3), 437–456. doi:10.1037/0033-2909.121.3.437

Whissell, C. M. (1989). The dictionary of affect in language. In R. Plutchik & H. Kellerman (Ed.), *Emotion: Theory, research and experience. The measurement of emotions* (Vol. 4, pp. 113-131). New York: Academic Press.

Wierzbicka, A. (1992). Talking about emotions: Semantics, culture, and cognition. *Cognition and Emotion*, *6*, 3–4. doi:10.1080/02699939208411073

Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., et al. (2008, September). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Proceedings of Interspeech,* Brisbane, Australia (pp. 597-600).

Wu, L., Oviatt, S. L., & Cohen, P. R. (1999). Multimodal integration: A statistical view. *IEEE Transactions on Multimedia*, *1*(4), 334–341. doi:10.1109/6046.807953

Xu, M., Jin, J. S., Luo, S., & Duan, L. (2008, October). Hierarchical movie affective content analysis based on arousal and valence features. In *Proceedings of ACM Multimedia,* Vancouver, British Columbia, Canada (pp. 677-680). ACM Publishing.

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2007, July). Music emotion classification: A regression approach. In *Proceedings of the IEEE International Conference on Multimedia and Expo,* Beijing, China (pp. 208-211). Washington, DC: IEEE Computer Society.

Yilmaz, A., Javed, O., & Shah, M. (2006). Object tracking: A survey. *ACM Journal of Computing Surveys*, *38*(4), 1–45.

Yu, C., Aoki, P. M., & Woodruff, A. (2004, October). Detecting user engagement in everyday conversations. In *Proceedings of 8th International Conference on Spoken Language Processing,* Jeju Island, Korea (pp. 1329-1332).

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. doi:10.1109/TPAMI.2008.52

Zhang, S., Tian, Q., Jiang, S., Huang, Q., & Gao, W. (2008, June). Affective MTV analysis based on arousal and valence features. In *Proceedings of the IEEE International Conference on Multimedia and Expo,* Hannover, Germany (pp. 1369-1372). Washington, DC: IEEE Computer Society.

Hatice Gunes received the PhD degree in computing sciences from the University of Technology Sydney (UTS), Australia, in 2007. From 2006 to 2008, she was a postdoctoral research associate at UTS, where she worked on an Australian research council–funded Linkage Project for UTS and iOmniscient Pty Ltd. She is currently a postdoctoral research associate at the HCI2 Group, Visual Information Processing Section, Imperial College London, U.K. working on a European Commission (EC-FP7) project, and is also an honorary associate of UTS. Gunes has published over 30 technical papers in the areas of video analysis and pattern recognition, with applications to video surveillance, human-computer interaction, emotion recognition and affective computing. She is a member of the IEEE and the ACM.

Maja Pantic received her MSc and PhD degrees in computer science from Delft University of Technology, the Netherlands, in 1997 and 2001, respectively. Until 2006, she was an assistant and then an associate professor at the same university. In 2006, she joined Imperial College London, UK, Computing Department, where she is reader in Multimodal HCI, and University

of Twente, the Netherlands, Department of Computer Science, where she is professor of affective and behavioral computing. She is the editor in chief of Image and Vision Computing Journal and an associate editor for the IEEE Transactions on Systems, Man, and Cybernetics, Part B. In 2008, she received European research council starting grant, as one of 2.5% of the best young researchers in Europe in any scientific discipline. Her research interests include computer vision and machine learning applied to face and body gesture recognition, multimodal HCI, and affective computing. She has published more than 90 technical papers in these areas and she has more than 1500 citations to her work. She is a senior member of the IEEE and a member of the ACM.