10

From the Lab to the Real World: Affect Recognition Using Multiple Cues and Modalities

Hatice Gunes[1], Massimo Piccardi[1] and Maja Pantic[2,3]
[1]*University of Technology, Sydney (UTS)*
[2]*Imperial College, London, United Kingdom*
[3]*University of Twente, the Netherlands*
[1]*Australia*
[2]*United Kingdom*
[3] *The Netherlands*

1. Introduction

Human affect sensing can be obtained from a broad range of behavioral cues and signals that are available via visual, acoustic, and tactual expressions or presentations of emotions. Affective states can thus be recognized from visible/external signals such as gestures (e.g., facial expressions, body gestures, head movements, etc.), and speech (e.g., parameters such as pitch, energy, frequency and duration), or invisible/internal signals such as physiological signals (e.g., heart rate, skin conductivity, salivation, etc.), brain and scalp signals, and thermal infrared imagery.

Despite the available range of cues and modalities in human-human interaction (HHI), the mainstream emotion research has mostly focused on facial expressions (Hadjikhani & De Gelder, 2003). In line with this, most of the past research on affect sensing and recognition has also focused on facial expressions and on data that has been posed on demand or acquired in laboratory settings. Additionally, each sense such as vision, hearing, and touch has been considered in isolation. However, natural human-human interaction is multimodal and not occurring in predetermined, restricted and controlled settings. In the day-to-day world people do not present themselves to others as voice- or body-less faces or face- or body-less voices (Walker-Andrews, 1997). Moreover, the available emotional signals such as facial expression, head movement, hand gestures, and voice are unified in space and time (see Figure 1). They inherently share the same spatial location, and their occurrences are temporally synchronized. Cognitive neuroscience research thus claims that information coming from various modalities is combined in our brains to yield multimodally determined percepts (Driver & Spence, 2000). In real life situations, our different senses receive correlated information about the same external event. When assessing each others' emotional or affective state, we are capable of handling significantly variable conditions in terms of viewpoint (i.e. frontal, profile, even back view), tilt angle, distance (i.e., face to face as well as at a distance) , illumination (i.e., both day and night conditions), occlusions (e.g., even when some body parts are occluded), motion (e.g., both when stationary and moving, walking and talking) and noise (e.g., while many people are chatting and interacting simultaneously).

The fact that humans perceive the world using rather complex multimodal systems does not necessarily imply that the machines should also posses all of the aforementioned functionalities. Humans need to operate in all possible situations and develop an adaptive behavior; machines instead can be highly profiled for a specific purpose, scenario, user, etc. For example, the computer inside an automatic teller machine probably does not need to recognize the affective states of a human. However, in other applications (e.g., computer agents, effective tutoring systems, clinical settings, monitoring user's stress level) where computers take on a social role such as an *instructor* or *helper,* recognizing users' affective states may enhance the computers' functionality (Picard, 1997).

A number of survey papers exist within the affect sensing and recognition literature (e.g., Gunes & Piccardi, 2008; Zeng & et al., 2008). For instance, the shift from monomodal to multimodal affect recognition, together with systems using vision as one of the input modalities and analyzing affective face and body movement either as a pure monomodal system or as part of a multimodal affective framework, is discussed in (Gunes & Piccardi, 2008). An exhaustive survey of past efforts in audiovisual affect sensing and recognition, together with various visual, audio and audio-visual databases, is presented in (Zeng & et al., 2008). However, no effort so far has attempted to compile and discuss visual (i.e., facial and bodily expression), audio, tactile (i.e., heart rate, skin conductivity, thermal signals etc.) and thought (i.e., brain and scalp signals) modalities together. Accordingly, this chapter sets out to explore recent advances in affect sensing and recognition by explicitly focusing on systems that are based on multiple input modalities and alternative channels, and is organized as follows. The first part is concerned with the challenges faced when moving from affect recognition systems that were designed in and for laboratory settings (i.e., analyzing posed data) to systems that are able to analyze spontaneous data in a multimodal framework. It discusses the problem domain of multimodal affect sensing, when moving from posed to spontaneous settings. The chapter initially focuses on background research, reviewing the theories of emotion, monomodal expression and perception of emotions, temporal information, posed vs. spontaneous expressions, and multimodal expression and perception of emotions. The chapter then explores further issues in data acquisition, data annotation, feature extraction, and multimodal affective state recognition. As affect recognition systems using multiple cues and modalities have only recently emerged, the next part of the chapter presents representative systems introduced during the period 2004 - 2007, based on multiple visual cues (i.e., affective head, face and/or body movement), haptic cues (physiological sensing) or combination of modalities (i.e., visual and physiological channels, etc.) capable of handling data acquired either in the laboratory or real world settings. There exist some studies analyzing spontaneous facial expression data in the context of cognitive-science or medical applications (e.g., Ashraf & et al., 2007). However, the focus of this chapter is on multimodal or multicue affective data, accordingly, systems analyzing spontaneous data are presented in the context of human-computer interaction (HCI) and human-robot interaction (HRI). The last part of this chapter discusses issues to be explored in order to advance the state-of-the-art in multimodal and multicue affect sensing and recognition.

2. From posed to spontaneous: changes and challenges

Affect sensing and recognition is a relatively new research field. However, it should be realized that affect recognition from multiple modalities has an even shorter historical background and is still in its infancy. It was not till 1998 that computer scientists attempted to use multiple modalities for recognition of emotions/affective states (Riseberg & et al., 1998). The initial interest was on fusing visual and audio data. The results were promising; using multiple modalities improved the overall recognition accuracy helping the systems function in a more efficient and reliable way. Starting from the well-known work of Picard (Picard & et al., 2001), interest in detecting emotions from physiological signals emerged. Although a fundamental study by Ambady and Rosenthal suggested that the most significant channels for judging behavioral cues of humans appear to be the visual channels of face and body (Ambady & Rosenthal, 1992), the existing literature on automatic emotion recognition did not focus on the expressive information that body gestures carry till 2003 (e.g., Camurri & et al., 2003). Following the new findings in psychology, a number of researchers have attempted to combine facial expressions and body gestures for affect recognition (e.g., Gunes & Piccardi, 2007; Karpouzis & et al., 2007; Martin & et al., 2006). A number of approaches have also been proposed for other sensorial sources such as thermal and brain signals (e.g., Nakasone & et al., 2005; Takahashi, 2004; Pun & et al., 2006; Puri & et al., 2005; Savran & et al., 2006; Takahashi, 2004; Tsiamyrtzis & et al., 2007). With all these new areas of research in affect sensing, a number of challenges have arisen (e.g., synchronization, fusion, etc.). The stage that affective computing has reached today is combining multiple channels for affect recognition and moving from laboratory settings towards real world settings.

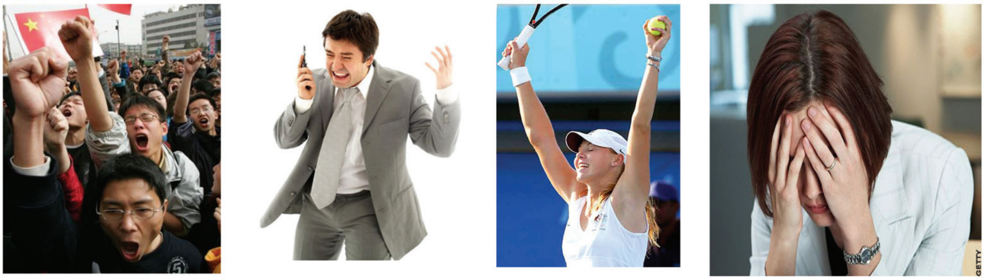


Figure 1: Examples of socially visible multimodal expression (facial expression, body gesture and speech) of emotions in real-life situations.

We start with the description of what is meant by *laboratory* vs. *real world settings*. The so-called laboratory/posed/controlled settings refer to:

- an experimental setup or environment (e.g., a laboratory), with controlled and uniform background/illumination/placement conditions (e.g., a static background without any clutter, no audiovisual noise, with predetermined level of illumination and number of lights etc.),
- human subject restricted in terms of free movement of head/body and in terms of location/seating and expressivity s(he) is allowed/able to display,
- a setup where people are instructed by an experimenter on how to show the desired actions/expressions (e.g., by moving the left eyebrow up or producing a smile), where occurrences of occlusion/noise/missing data are not allowed,
- a setup without considering any of the issues related to user, task or context.

The so-called real world/spontaneous/natural settings instead refer to:

- a realistic environment, for instance, home/office/hospital, without attempting to control the varying conditions,

- where people might show all possible affective states, expressed synchronously (e.g., speech and facial expression) or asynchronously (e.g., facial expression and body gesticulation), expressed with intention (e.g., irony) or without intention (e.g., fatigue),
- with large head or body movements as well as moving subjects in various environments (e.g., office or house, not just restricted to one chair or room),
- where people are not aware of the recording (or are, depending on the context),
- where people will not restrain themselves unlike the case when they are part of an experiment, and will express emotions due to real-life event or trigger of events (e.g., stressed at work),
- with possible occurrences of occlusions (e.g., hands occluding each other or hand occluding the face), noise (e.g., in audio recordings) and missing data,
- where the recordings are acquired with multiple sensing devices (e.g., multiple cameras & microphones & haptic/olfactory/taste/brain sensors etc.), under non-uniform and noisy (lighting/voice recording) conditions and in long sessions (e.g., one whole day and possibly a couple of weeks or longer),
- capturing all variations of expressive behavior in every possible order/combination/scale,
- being able to adapt to user, task and context.

As the real world settings pose many challenges to the automatic sensing and recognition of human affect, there have been a relatively higher number of research studies on affect recognition that have dealt with laboratory settings rather than real world settings. The shift from the laboratory to the real world is driven by various advances and demands, and funded by various research projects (e.g., European Union FP 6, HUMAINE and European Union FP 7, SEMAINE). However, similar to that of many other research fields, the shift is gradual and the progress is slow. The multimodal systems introduced so far can only partially handle challenges mentioned as part of the more naturalistic or real world settings. Although multimodal systems or machines aimed at assisting human users in their tasks might not need to function exactly as humans do, it is still necessary to investigate which modalities are the most suitable ones for which application context. To date, many research questions remain unexplored while advancing toward that goal.

3. Background research

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

3.1 Theories of emotion

One of the most discussed issues in the emotion literature is the definition, categorization and elicitation of emotions. As far as definition of emotion is concerned emotions are defined as affectively valenced states (Ortony & Turner, 1990). In general, emotions are short-term (seconds/minutes), whereas moods are long-term (several days), and temperaments or personalities are very long-term (months, years or a lifetime) (Jenkis & et al., 1998).

As far as the categorization is concerned, a significant number of researchers in psychology advocate the idea that there exists a small number of emotions that are basic as they are hard-wired to our brain and are recognized universally (e.g., (Ekman & et al., 2003). Ekman and his colleagues conducted various experiments on human judgment on still photographs of posed facial behavior and concluded that the six basic emotions can be recognized universally, namely, happiness, sadness, surprise, fear, anger and disgust (Ekman, 1982). To date, Ekman's theory on universality is the most widely used theory in affect sensing by machines.

Some other researchers argue about how many emotions are basic, which emotions are basic, and why they are basic (Ortony & Turner, 1990). Some researchers claim that the list of basic emotions (i.e., happiness, surprise, desire, fear, love, rage, sadness etc.) includes words that do not refer to emotions. For instance, a few researchers claim that surprise is an affectively neutral state; therefore is not an emotion (Ortony & Turner, 1990).

Among the various classification schemes, Baron-Cohen and his colleagues, for instance, have investigated cognitive mental states (e.g., agreement, concentrating, disagreement, thinking, unsure and interested) and their use (see Figure 2a) in daily life 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 day to day interactions than the so-called basic emotions (Baron-Cohen & Tead, 2003). These states were also found relevant in representing problem-solving and decision-making processes in HCI context and have been used by a number of researchers (e.g., El Kaliouby & Robinson, 2005).

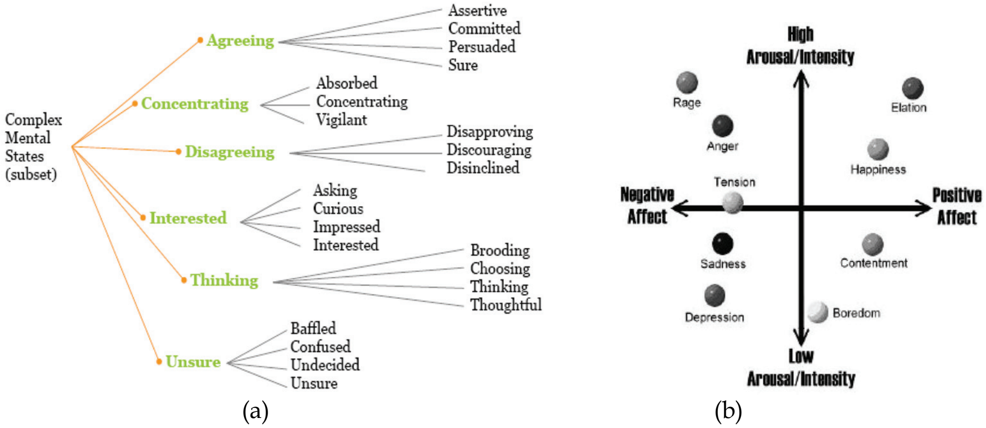


Figure 2. Illustration of a) Baron-Cohen's cognitive mental states (from Autism and Affective-Social Computing Tutorial at ACII 2007), and b) Russell's circumflex model (Russell, 1980).

A number of emotion researchers take the dimensional approach and they view affective states not independent of one another; rather, related to one another in a systematic manner (e.g., Russell, 1980). Russell (Russell, 1980) among others argues that emotion is best characterized in terms of a small number of latent dimensions, rather than in terms of a small number of discrete emotion categories. Russell proposes that each of the basic emotions is a bipolar entity as part of the same emotional continuum. The proposed polars are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). Arousal is a feeling state that ranges from sleepiness or boredom to frantic excitement. Valence ranges from

unpleasant feelings to pleasant feelings of happiness. The model is illustrated in Figure 2b.
Another issue in the emotion research is that of certain emotions' co-occurrence. Russell and Carroll (Russell & Carroll, 1999), in accordance with Russell's circumflex model, propose that happiness and sadness are polar opposites and, thus, mutually exclusive. In other words, “when you are happy, you are not sad and when you are sad, you are not happy". In contrast, Cacioppo and Berntson (Cacioppo & Berntson, 1994) propose that positive and negative affect are separable, and mixed feelings of happiness and sadness can co-occur.

As far as the labeling is concerned, different labels are used by different researchers when referring to the same emotion (e.g., anger - rage, happiness - joy - elation). The problem of what different emotion words are used to refer to the same emotion is by itself a topic of research for linguists, emotion theorists, psychologists and potentially ethnologists (Ortony & Turner, 1990).

After all, even with over a century of research, all of the aforementioned issues still remain under discussion and psychologists do not seem to have reached consensus yet. In relevance to this chapter, in the following sections, background in nonverbal communication of emotions is provided. In particular, studies that explore the characteristic nonverbal expressions of emotions in HHI from various channels are reviewed under two categories: i) monomodal expression and perception of emotions and ii) multimodal expression and perception of emotions.

3.2 Monomodal expression and perception of emotions

Emotional information is conveyed by a broad range of modalities, including speech and language, gesture and head movement, body movement and posture, as well as facial expression. One limitation of prior work on human emotion perception is the focus on separate channels for expression of affect, without adequate consideration for the multimodal emotional signals that people encounter in their environment (Ekman, 1982; Pantic & Rothkrantz, 2003; Van den Stock & et al., 2007). Most research on the development of emotion perception has focused on human recognition of facial expressions and posed emotional data. The investigation of various ways in which people learn to perceive and attend to emotions multimodally is likely to provide a more complete picture of the complex HHI.

Herewith, we provide a summary of the findings from emotion research in emotion communication from facial and bodily expression, audio or acoustic signals, and bio-potential signals (physiological signals, brain signals and thermal infrared signals). Figure 3 presents examples of sensors used for acquiring affective data from these channels.

3.2.1 Facial expression

Ekman and his colleagues conducted various experiments on human judgment on still photographs of posed face behavior 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 but it remains unconfirmed whether they are universally distinguishable. Although prototypic expressions, like happiness, surprise and fear, are natural, they occur infrequently in daily life and provide an incomplete description of facial expression. To capture the subtlety of human emotion and paralinguistic communication, Ekman and Friesen developed the Facial Action Coding System (FACS) for coding of fine-grained changes on the face (Ekman &

Friesen, 1978). FACS is based on the enumeration of all *face action units* causing face movements. In addition to this, Friesen and Ekman (Friesen & Ekman ,1984) developed Emotion FACS (EMFACS) as a method for using FACS to score only the facial actions that might be relevant to detecting emotion.

To date, Ekman's theory of emotion universality (Ekman & Friesen, 2003) and the Facial Action Coding System (FACS) (Ekman &Friesen, 1978) are the most commonly used schemes in vision-based systems attempting to recognize facial expressions and action units.

3.2.2 Bodily expression

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

Darwin (Darwin, 1872) 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. It is also well known from animal research that information from bodily expressions can play a role in reducing the ambiguity of facial expression (Van Hoof, 1962). It has been shown that observers' judgments of infant emotional states depend on viewing whole-body behaviors more than facial expression (Camras & et al., 2002). 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 & et al. (Van den Stock & et al., 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 sad bodily expressions. (Gross & et al., 2007) presented a study where bodily expression of felt and recognized emotions was associated with emotion specific changes in gait parameters and kinematics (content, joy, angry, sad and neutral). After recalling an emotion, participants walked across the laboratory while video and whole-body motion capture data were acquired. Walkers felt and observers recognized the same emotion in 67% of the available data. On average, sadness was most recognized and anger was least recognized. Gait velocity was greatest in high-activation emotion trials (anger and joy), and least in sad trials. Velocity was not different among neutral and low-activation emotion trials (content and sad). Both posture and limb motions changed with emotion expressed.

In general, the body and hand gestures are much more varied than face gestures. There is an unlimited vocabulary of body postures and gestures with combinations of movements of various body parts. Despite the effort of Laban in analyzing and annotating body movement (Laban & Ullmann, 1988) unlike the face action units, body action units that carry expressive information have not been defined or coded with a Body Action Coding System. 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 the various affective states.

3.2.3 Audio

Speech is another important communicative modality in human-human interaction. It is between 200 thousand and 2 million years old, and it has become the indispensable means for sharing ideas, observations, and feelings. 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 non-verbal aspect of the speech. However, findings in basic research indicate that spoken messages are rather unreliable means to analyze and predict human (affective) behavior (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.

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 & Scherer (2005). The speech measures 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 nonbasic affective states such as distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, and yawns (Russell & Fernandez-Dols, 1997).

3.2.4 Bio-potential signals

Brain signals measured via functional Near Infrared Spectroscopy (fNIRS), scalp signals measured via electroencephalogram (EEG), and peripheral signals, namely, cardiovascular activity, including interbeat interval, relative pulse volume, pulse transit time, heart sound, and pre-ejection period; electrodermal activity (tonic and phasic response from skin conductance) or galvanic skin response (GSR), electromyogram (EMG) activity (from corrugator supercilii, zygomaticus, and upper trapezius muscles), are commonly referred to as physiological or bio-signals (Changchun & et al., 2005; Savran & et al., 2006; Takashi, 2004). While visual modalities such as facial expressions and body gestures provide a visible/external understanding of the emotions, bio-signals such as EEG and fNIRS provide an invisible/internal understanding of the emotion phenomenon (see (Savran & et al., 2006) and Figure 3).

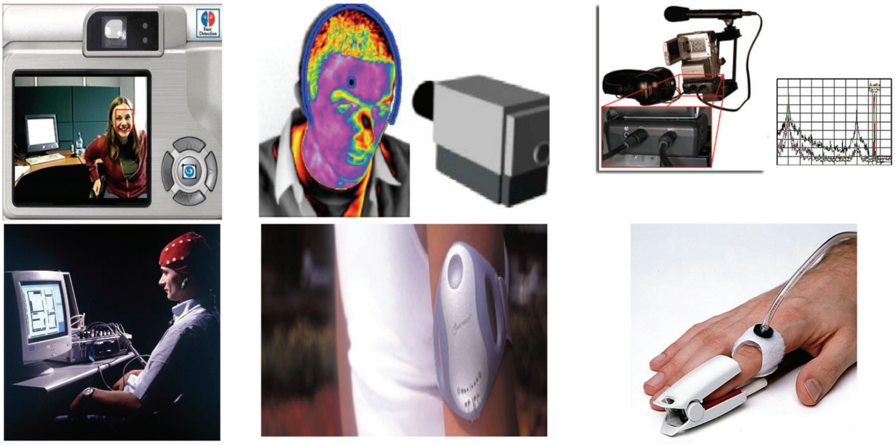


Figure 3: Examples of sensors used in multimodal affective data acquisition: (a) camera for visible imagery, (b) microphone(s) for audio recording, (c) infrared camera for thermal infrared (IR) imagery, (d) body media sense wear for physiological signal recording, (e) pulse wave signal recorder clipped on the finger, and (f) electroencephalogram (EEG) for brain/scalp signals recording and measurement.

Researchers claim that all emotions can be characterized in terms of judged valence (pleasant or unpleasant) and arousal (calm or aroused) (Lang, 1995). Emotions can thus be represented as coordinates in the arousal–valence space. The relation between physiological signals and arousal/valence is established in psychophysiology that argues that the activation of the autonomic nervous system changes while emotions are elicited (Levenson, 1988). Galvanic skin response (GSR) is an indicator of skin conductance (SC), and increases linearly with a person's level of overall arousal and Electromyography (EMG) measures muscle activity and has been shown to correlate with negatively valenced emotions (Nakasone & et al., 2005). The transition from one emotional state to another, for instance, from state of boredom to state of anxiety is accompanied by dynamic shifts in indicators of autonomic nervous system activity (Changchun & et al., 2005). Moreover, 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. Therefore, such responses can be used as correspondence to positive/negative emotions (Pun & et al., 2006). BCIs can assess the emotions by assuming that negative/positive valence corresponds to negative/ positive emotions and arousal corresponds to the degree of excitation, from none to high (e.g., (Pun & et al., 2006)). However, in general, researchers have not identified an optimal set of bio-potential cues that can assist in reliably discriminating among various affective states.

3.2.5 Thermal infrared signals

A number of studies in the fields of neuropsychology, physiology and behavior analysis suggest that there exists a correlation between mammals' core body temperature and their affective states. Nakayama & et al. conducted experiments by monitoring the facial temperature change of monkeys under stressful or threatening conditions. Their study revealed that decrease in nasal skin temperature is relevant to a change from neutral to negative affective state (Nakayama & et al., 2005). Vianna & Carrive conducted another independent experiment by monitoring the temperature changes in rats when they were experiencing fearful situations (Vianna & Carrive, 2005). The observation was 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.

There also exist other studies indicating that contraction or expansion of the facial/bodily muscles of humans causes fluctuations in the rate of blood flow (e.g., Khan & et al., 2006a, 2006b; Tsiamyrtzis & et al., 2007). This thermo-muscular activity results in a change in the volume of blood flow under the surface of the human facial and/or bodily skin. Thus, skin temperature is heavily modulated by superficial blood flow and environmental temperature. However, influence of environmental temperature blends in the background once the person is in that environment and can be modeled or ignored. This in turn implies that it is possible to obtain objective measurements of skin temperature change.

Unlike other bio-physiological signal measurement, sensing using infrared thermal imagery does not rely on contact with the human body. Thus, the noninvasive detection of any changes in facial and/or bodily thermal features may help in detecting, extracting, and interpreting human affective states. For instance, (Pavlidis & et al., 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, respectively. Similarly, Puri & et al. reported that users' stress level was correlated with increased blood flow in the frontal vessels of forehead causing dissipation of convective heat (Puri & et al., 2005).

A generic model for estimating the relationship between fluctuations in blood flow and facial/bodily muscle activity is not yet available. Such a model could enhance our understanding of the relationship between affective states and the facial/bodily thermal and physiological characteristics.

3.3 Temporal information

Studies show that the temporal dynamics play an important role for interpreting emotional displays (Ambady & Rosenthal, 1993; Schmidt & Cohn, 2001). The temporal aspect of a facial movement is described by four segments: neutral, onset, apex and offset (Ekman, 1979). The neutral phase is a plateau where there are no signs of muscular activation, the face is relaxed. The onset of the action/movement is when the muscular contraction begins and increases in intensity and the appearance of the face changes. The apex is a plateau usually where the intensity reaches a stable level and there are no more changes in facial appearance. The offset is the relaxation of the muscular action. A natural facial movement evolves over time in the following order: neutral- onset- apex-offsetneutral. Other combinations such as multiple-apex facial actions are also possible.

Similarly, the temporal structure of a body gesture consists of (up to) five phases: preparation (pre-stroke)- hold- stroke- (post-stroke) hold-retraction. The preparation moves to the stroke's starting position and the stroke is the most energetic part of the gesture. Holds are optional still phases which can occur before and/or after the stroke. The retraction returns to a rest pose (e.g., arms hanging down, resting in lap, or arms folded) (Wilson & et al., 1997).

As stated previously, research on bodily expression of emotions is relatively new. Moreover, most of the present studies on bodily expression of emotion have used static images, in line with the large majority of studies on facial expressions. Due to such reasons issues such as the importance of motion, timing, and spontaneity have not been considered as extensively as in the facial expression literature.

The importance of temporal information has also not been widely explored for bio-potential signals. Overall, similar body of research to the facial expressions needs to be conducted in order to identify the importance of such factors for bodily or bio-potential signal-based expressions of emotions and correlation between these cues and modalities. After all, detection of the temporal phases and/or dynamics can effectively support automated recognition of affective states (e.g., Gunes, 2007).

3.4 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 both deliberate/posed and natural/spontaneous emotional expressions are recognized equally accurately; however, deliberate expressions are significantly different from natural ones. Deliberate facial behavior is mediated by separate motor pathways and differences between natural and deliberate facial actions may be significant. Schmidt and Cohn (Schmidt & Cohn, 2001) found that an important visual cue signaling a smile as deliberate or spontaneous is the timing of the phases. A major body of research has been conducted by Cohn and his colleagues in order to identify such differences for other facial expressions of emotions (Affect analysis group, 2008).

In natural situations, 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 contained in the signal of body movements should be less significant than the face, at least when it comes to discerning spontaneous from posed behavior. Ekman however, argued that people do not bother to censor their body movements in daily life and therefore, the body would be the *leakier* source (Ekman, 2003). Furthermore, research in nonverbal behavior and communication theory stated that truthful and deceptive behavior differ from each other in lack of head movement (Buller & et al., 1994) and lack of illustrating gestures which accompany speech (DePaulo, 2003).

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 as to the user's state of mind.

Overall, perceiving dynamics for spontaneous emotional face and body language and recognition of dynamic whole bodily expressions has not been studied extensively. In day-today life people express and communicate emotions multimodally. Research that study posed vs. spontaneous expressions in a multicue and/or multimodal context therefore is needed in order to obtain a better understanding of the natural communication of emotions in HHI to be later used in HCI.

3.5 Multimodal expression and perception of emotions

In noisy situations, humans depend on access to more than one modality, and this is when the nonverbal modalities come into play (Cassell, 1998). It has been shown that when speech is ambiguous or in a speech situation with some noise, listeners do rely on gestural cues (McNeill, 1985).

Cross-modal integration is known to occur during multi-sensory perception. Judgments for one modality are influenced by a second modality, even when the latter modality can provide no information about the judged property itself or increase ambiguity (Driver & Spence, 2000). A number of studies reported that facial expressions and emotional tone of voice or emotional prosody influence each other (De Gelder & et al.,1999; Massaro & Cohen, 2000).. In a study with static facial expressions and emotional spoken sentences, de Gelder and Vroomen observed a cross-modal influence of the affective information. Recognition of morphed vocal expressions was biased toward the simultaneously presented facial expression, even when the participants were instructed to ignore the visual stimuli. A follow up study suggested that this cross-modal integration of affective information takes place automatically, independent of attentional factors (Vroomen & et al., 2001). Van den Stock & et al. (Van den Stock & et al., 2007) investigated the influence of wholebody expressions of emotions on the recognition of facial and vocal expressions of emotion. The recognition of facial expression was strongly influenced by the bodily expression. This effect was a function of the ambiguity of the facial expression. In another experiment they found that the recognition of emotional tone of voice was similarly influenced by task irrelevant emotional body expressions. Taken together, the findings illustrate the importance of emotional whole-body expressions in communication when viewed in combination with facial expressions and emotional voices.

When input from multiple expressive sources or channels is available the affective message conveyed by different modalities might be congruent (i.e., agreeing) or incongruent (i.e., disagreeing). Observers judging a facial expression were found to be strongly influenced by emotional body language (Meeren & et al., 2005). (Meeren & et al., 2005) investigated the combined perception of human facial and bodily expressions. Participants were presented compound images of faces on bodies and their emotional content was either congruent or incongruent. The results showed that responses were more accurate and faster when face and body expressed the same emotion. When face and body convey conflicting emotional information, judgment of facial expression is hampered and becomes biased toward the emotion expressed by the body. The results show that the whole-body expression has the most influence when the face ambiguity is highest and decreases with reduced facial ambiguity.

Emotion research has not reported such cross-modal interaction for other pairs of modalities such as tactile and visual, or tactile and audio etc. These issues need to be addressed in follow-up studies to obtain a better understanding of the interaction between various expressive cues, sources and modalities in HHI. The multimodal affect systems should potentially be able to detect incongruent messages and label them as *incongruent* for further/detailed understanding of the information being conveyed (Paleari & Lisetti, 2006) .

Different to the cross-mode compensation but still part of the multicue or multimodal perception, there exist findings reporting that when distance is involved humans tend to process the overall global information rather than considering configurations of local regions. Researchers found that if a face is present at close range, especially the eyes are important, but when the distance increases, the configural properties of the whole face play an important role (Van den Stock & et al., 2007). Whole-body expressions seem to be preferentially processed when the perceiver is further away from the stimulus. When the facial expression of the producer is not visible, emotional body language becomes particularly important. Such issues are yet to be explored in multimodal affect recognition.

If humans are presented with temporally aligned but conflicting audio and visual stimuli, the perceived sound may differ from that present in either channel. This is known as McGurk effect in the audio-visual speech perception literature. (Ali & et al., 2003) examined the effect of temporal misalignment of audio and visual channels when users interact with multimodal interfaces (e.g., talking heads). Their study showed that when the audio is not in synchrony with the visual channel, the McGurk effect is observed and participants need to apply extra mental effort for recognition. Such an analysis has not yet been applied in the field of affect sensing and recognition.

Overall, further research is needed in multicue and multimodal affect sensing and recognition in order to explore the issues that have been discussed in this section.

4. Data acquisition

A recent discussion in the automatic affect sensing field is the creation and use of posed vs. spontaneous databases. Affective data may belong to one of the following categories: posed (i.e., produced by the subject upon request), induced (i.e., occurring in a controlled setting and designed to create an affective activation or response such as watching movies) or spontaneous (i.e., occurring in real-life settings such as interviews or interactions between humans or between humans and machines) (Bänziger and Scherer, 2007).

When acquiring posed affective multimodal data, the experiments are usually carried out in a laboratory setting where the illumination, sounds, and room temperature are controlled to maintain uniformity. The stimulated emotions usually include the so-called six basic emotions (e.g., Takashi, 2004). Posed databases are recorded by asking "actors" to act specific affective-cognitive states. The easiest way to create a posed multimodal affect database is by having an experimenter direct and control the expression/display and the recordings. The creation of such database usually depends on the restrictions imposed on the actors: e.g., where the subject should sit or stand, where the subject should look, how a smile should be displayed, whether or not head motion, body motion or speech are allowed etc. Moreover, transitions between affective states are not allowed. Depending on which modalities are recorded, the experimenters typically use a number of sensors: two cameras where face and upper body are recorded simultaneously (e.g., the FABO database (Gunes & Piccardi, 2006)), a camera and a microphone when recording facial expressions and audio signals (e.g., the University of Texas Database (O'Toole & et al., 2005)) etc. A typical affective state recorded thus consists of neutral-onset-apex-offset-neutral temporal segments. When acquiring spontaneous affective multimodal data, the subjects are recorded without their knowledge while they are stimulated with some emotionally-rich stimulus (e.g., Zuckerman & et al., 1979). In the facial expression recognition literature the so-called spontaneous data is facial behavior in less constrained conditions such as an interview setting where subjects are still aware of placement of cameras and their locations (e.g., Littlewort & et al., 2007; Pantic & Bartlett (2007).

Recording of the physiological or 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 on which electrodes are mounted, a clip sensor, and/or touch type electrodes. The subject is then stimulated with emotionally-evocative images/videos/sounds. EEG recordings capture neural electrical activity on a millisecond scale from the entire cortical surface while fNIRS records hemodynamic reactions to neural signals on a seconds scale from the frontal lobe. The skin conductance meter is usually composed of a number of electrodes and an amplifier. The electrodes are mounted on a surface, for example a mouse in order to contact the fingers of the subject. The variation of the skin conductance at the region of interest is then measured (Takahashi, 2004). In summary, the bio-potential affect data acquisition is *induced*, however, due to its invasive nature, the experimental settings provided do not encourage spontaneity.

Recently, there have been a number of attempts to create publicly available affect databases using multiple sensors or input modalities. Some examples can be listed as follows: the SmartKom Corpora, the FABO Database, the Database collected at the University of Amsterdam and the Database collected at the University of Texas. These databases have been reviewed in (Gunes & Piccardi, 2006b) in detail. Such affect databases fall in either the posed or induced category. A number of databases (e.g.,: Drivawork Database, SAL Database and Mixed/spaghetti Data) have also been created as part of HUMAINE EU FP6 and have been presented in (Douglas-Cowie & et al., 2007). Among these, the Belfast database is a naturalistic audio-visual database consisting of clips taken from television and realistic interviews with a research team, and the SAL database contains induced data where subjects interacted with artificial listener with different personalities were recorded audio-visually.

Creation and annotation of affect databases from face and body display has been reviewed in (Gunes & Piccardi, 2006b). Various visual, audio and audio-visual databases have been reviewed in (Zeng & et al., 2008).

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 the context of autism disorder) (El Kaliouby & Teeters, 2007), such an attempt is lacking for multimodal affect database creation. Overall, acquiring data in fully unconstrained environments with multiple sensors involves ethical and privacy concerns together with technical difficulties (placement of sensors, controlling the environmental conditions such as noise, illumination, occlusions, etc., consistency, repeatability etc.).

5. Data annotation

The relative weight given to facial expression, speech, and body cues depend both on the judgment task (i.e. what is rated and labeled) and the conditions in which the behavior occurred (i.e. how the subjects were simulated to produce the expression) (Ekman, 1982). People do not judge the available communicative channels separately and the information conveyed by these channels cannot be assumed simply additive (i.e., cross-mode compensation). However, in general, annotation of the data in multimodal affect databases, both for posed and spontaneous data, has been done separately for each channel assuming independency between the channels.

The experimental setup for labeling or annotating emotional behaviors expressed via the visual channels usually consist of static photographs (e.g., Van den Stock & et al., 2007) or videos containing semi-professional actors expressing six basic (or more) emotions with face (e.g., Bänziger & Scherer, 2007), face and upper body (e.g., Gunes & Piccardi, 2006a), whole-body with faces blurred (e.g., Van den Stock & et al., 2007), or stick figures (e.g., Coulson, 2004). Visual data are presented on a computer screen, and participants are asked to view and choose an emotion label from a predetermined list of labels that best describes the expressed emotion. Such studies usually aim to determine rates of observer recognition in visual stimuli, and to use motion analysis to quantify how emotions change patterns in characteristic ways.

In general, when annotating or labeling affect data from face display, Ekman's theory of emotion universality and the Facial Action Coding System (FACS) are used. When it comes to annotating body gestures, unlike the AUs, there is not one common annotation scheme that can be adopted by all the research groups. Laban and Ullman defined and analyzed body movement by using the following information and descriptions: body part (e.g., left hand), direction (e.g., up/down), speed (e.g., fast/slow), shape (hands made into fists), space (flexible/direct), weight (light/strong), time (sustained/quick), and flow (fluent/controlled) (Laban & Ullmann, 1988). Overall, the most time-costly aspect of current facial/body movement manual annotation is to obtain the onset-apex-offset time markers. This information is crucial for coordinating facial/body activity with simultaneous changes in physiology, voice, or speech.

Hereby we describe some attempts or the so-called *coding schemes* for annotating multimodal affect data. In (Allwood & et al., 2004) authors designed a coding scheme for the annotation of 3 videos of TV interviews. Facial displays, gestures, and speech were coded using the following parameters: form of the expression and of its semantic-pragmatic function (e.g. turn managing) and the relation between modalities: repetition, addition, substitution, contradiction. (Martin & et al., 2005) designed a coding scheme for annotating multimodal behaviors during real life mixed emotions (i.e., TV interviews). They focused on the annotation of emotion specific behaviors in speech, head and torso movements, facial expressions, gaze, and hand gestures. They grounded their coding scheme on the following parameters: the expressivity of movements, the number of annotations in each modality, their temporal features (duration, alternation, repetition, and structural descriptions of gestures), the directions of movements and the functional description of relevant gestures. (Martin & et al., 2007) designed a multilevel coding scheme for the representation of emotion at several levels of temporality and abstraction. At the global level there is the annotation of emotion (categorical and dimensional including global activation). Similar annotations are available at the level of emotional segments of the video. At the level of multimodal behaviors there are tracks for each visible behavioral cue: torso, head, shoulders, facial expressions, gaze, and hand gestures. The head, torso and hand tracks contain a description of the pose and the movement of these cues. For the annotation of emotional movements, they use a model which describes expressivity by a set of six dimensions: spatial extent, temporal extent, power, fluidity, repetition, overall activity. The annotation also includes the structural descriptions (phases) of gestures.

When annotating or labeling affect data from audio participants are asked to identify an emotion (e.g., happy or sad) given an auditory spoken sentence. Thus, again Ekman's theory of emotion universality or Russell's theory of arousal and valence is the most common way to annotate audio signals.

For bio-potential signal annotation, ground truth usually consists of the participant’s self-assessment (e.g., Pun & et al., 2006). In general, Ekman's theory of emotion universality or Russell's theory of arousal and valence is used. However, obtaining a correlation between emotions and the neural, thermal and other signals is not a straightforward process and is inherently different compared to visual or audio channels. The data labeling for bio-signals is directly dependant on the individuals’ evaluation of his own emotional situation during the experimental setup (i.e., emotion elicitation) or recordings. This implies that, the ground truth coding or labeling is very subjective and cannot be evaluated by independent observers or emotion research experts.

Another major challenge in affect data annotation is the fact that there is no coding scheme that is agreed upon by all the researchers to accommodate all possible communicative modalities like facial expressions, body gestures, voice, bio-potential signals etc. Addressing the aforementioned issues will potentially extend the state-of-the-art in multimodal affect sensing and recognition.

6. Feature extraction

After multimodal affect data has been acquired and annotated, representative and relevant features need to be extracted prior to the automatic affect recognition procedure. The feature extraction is only broadly covered here under each communicative source or channel: facial expression, body gestures, audio, bio-potential signals and thermal infrared imagery.

6.1 Facial expression

There exists an extensive literature for human face detection, feature extraction and expression recognition. Possible strategies for face detection vary significantly depending on the type of input images. Face detection can become a simplified problem with the assumption that an input image contains only one face. The so-called appearance-based methods have proved very robust and fast in recent years. They usually are based on training a classifier using positive and negative examples of faces. Various classifiers can be used for this purpose: Naive Bayes classifier, Hidden Markov model (HMM), Sparse network of windows (SNoW), Support Vector Machines (SVM), Adaboost etc. For face detection, the current state-of-the-art is based on the robust and well-known method proposed by Viola and Jones (Viola & Jones, 2001) and extended by Lienhart and Maydt based on a set of rotated Haar-like features (Lienhart & Maydt, 2002), and improved by (Fasel & et al., 2005) using GentleBoost.

Facial feature extraction aims to detect the presence and location of features such as eyes, nose, nostrils, eyebrow, mouth, lips, ears, etc. Similar to face detection, for facial feature extraction usually it is assumed that there is only one face in the image. There exists an extensive literature for face feature extraction for the detection of facial region and facial features using texture, depth, shape, color information or statistical models. Such approaches can be categorized under two categories: feature-based approaches and appearance-based approaches. In the feature-based approach, specific facial features such as the pupils, inner/outer corners of the eyes/ mouth are detected and tracked, distances between these are measured or used and prior knowledge about the facial anatomy is utilized. In the appearance-based approach, certain regions are treated as a whole and motion, change in texture are measured. A similar approach to face detection can also be used for training a separate classifier on each facial feature (eyes, lips etc.). Such an approach can handle inplane rotation and tolerate variations in lighting. Methods based on Haar features or wavelets, also known as appearance-based methods, in general have demonstrated good empirical results. They are fast and fairly robust and can be easily extended to detect objects in different pose and orientation.

(Tian & et al., 2002) have shown that appearance-based methods alone perform poorly for the facial expression recognition. They found that when image sequences include nonhomogeneous subjects with small head motions, appearance-based methods have a relatively poor recognition rate compared to using an approach based on the geometric features and locations. Combining the two approaches (appearance based methods and geometric features) resulted in the best performance (Tian & et al., 2002). On the other hand, Bartlett and her colleagues have shown that appearance-based methods perform better than feature-based methods (Pantic & Bartlett, 2007). For further details on facial feature extraction and tracking for facial expression or action unit recognition the reader is advised to see (Pantic & Bartlett, 2007).

6.2 Body gesture

There exists an extensive literature for body feature extraction, tracking and gesture recognition from video sequences. In the context of affect sensing, we only briefly summarize the existing trends in body gesture recognition.

The existing approaches for hand or body gesture recognition and analysis of human motion in general can be classified into three major categories: model-based (i.e., modeling the body parts or recovering three-dimensional configuration of articulated body parts), appearance-based (i.e., based on two dimensional information such as color/gray scale images or body silhouettes and edges), and motion-based (i.e., using directly the motion information without any structural information about the physical body) (Elgammal, 2003). In the aforementioned approaches, Dynamic Time Warping (DTW) or Hidden Markov Models (HMM) are typically used to handle the temporal properties of the gesture(s).

An overview of the various tasks involved in motion analysis of the human body such as motion analysis involving human body parts, tracking of human motion using single or multiple cameras from image sequences is presented in (Yilmaz & et al., 2006). The literature on visual interpretation of hand gestures mainly focuses on HCI rather than affect sensing. (Mitra & Acharya, 2007) provide a recent survey on gesture recognition, with particular emphasis on hand gestures and facial expressions. Applications involving hidden Markov models, particle filtering and condensation, finite-state machines, optical flow, skin color, and connectionist models are discussed in detail. (Poppe, 2007) also provides a recent survey on vision-based human motion analysis and discusses the characteristics of human motion analysis via modeling and estimation phases.

Vision based gesture recognition is a challenging task due to various complexities including segmentation ambiguity and the spatio–temporal variability involved. Gesture tracking needs to handle variations in the tracked object (i.e., shapes and sizes of hands/arms) illumination, background, noise and occlusions. Recognition requires spotting of the gesture (i.e., determining the start and end points of a meaningful gesture pattern from a continuous stream) and segmenting the relevant gesture. Hand gestures may occlude each other as they switch from one gesture to another. Moreover, there occur intermediate and transition motions that may or may not be part of the gesture, and the same gesture may dynamically vary in shape and duration even for the same gesturer. Color as a distinct feature has been widely used for representation and tracking of multiple objects in a scene. Several tracking methods have been used in the literature; amongst them, the Kalman filter, Condensation tracking, Mean-shift tracking, and Cam-shift tracking. (Dreuw & et al. , 2006), for instance, present a dynamic programming framework with the possibility to integrate multiple scoring functions e.g. eigenfaces, or arbitrary objects, and the possibility of tracking multiple objects at the same time. Various techniques for extracting and tracking specific features such as shoulders have also been proposed in the literature. (Ning & et al., 2006) introduce a system that can detect shoulder shrug by firstly using a face detector based on AdaBoost and then detecting shoulder positions by fitting a parabola to the nearby horizontal edges using weighted Hough Transform to recognize shrugs. There are also more recent works using different (or a combination of) tracking schemes depending on what they aim to track and recognize. An example system is that of Valstar & et al. (Valstar & et al.,2007) that uses a cylindrical head tracker to track the head motion, an auxiliary particle filtering to track the shoulders motion, and a particle filtering with factorized likelihood tracking scheme to track movements of facial salient points in an input video. Overall, most of the existing hand/body gesture recognizers work well in relatively constrained environments (e.g., assuming that the person is seated on a chair) with relatively small changes in terms of illumination, background, and occlusions (Pantic & et al., 2007).

Compared to automatic gesture analysis and recognition, affective body gesture recognition per se has not been widely explored. For recognition of emotions from body movement and gesture dynamics, some researchers propose to extract the whole-body silhouette and the hands of the subjects from the background (e.g., Villalba & et al., 2007). Different motion cues are then calculated: amount of motion computed with silhouette motion images, the degree of contraction and expansion of the body computed using the bounding region, velocity and acceleration computed based on the trajectory of the hands etc. However, despite the challenges pertaining in the field, advance tracking techniques need to be created and used to be able to track body parts such as torso, head, shoulders, and hands in real world settings.

6.3 Audio features

Most of the existing approaches to vocal affect recognition use acoustic features, particularly pitch-related measures (range, mean, median, and variability), intensity, and duration, based on the acoustic correlations for emotion expressions as summarized by Cowie & et al. (2001). In addition, and mostly because they proved very suitable for speaker identification task, spectral features (e.g., MFCC, cepstral features) have been used in many of the current studies on automatic vocal affect recognition. Various studies have shown that pitch and energy among these features contribute most to affect recognition (Zeng & et al, 2008). A few efforts have been also reported that use some alternative features such as voice-quality features (Campbell & Mokhtari, 2003) and speech disfluencies (e.g., filler and silence pauses; Devillers & et al., 2004).

However, with the research shift towards analysis of spontaneous human behavior, it became clear that analysis of acoustic information only will not suffice for identifying subtle changes in vocal expression (Batliner & et al., 2003). In turn, several recent studies investigated the combination of acoustic features and linguistic features (language and discourse) to improve recognition of emotions from speech signal (e.g., Fragopanagos & Taylor, 2005). Examples include using spoken words and acoustic features, using prosodic features, spoken words, and information of repetition, and using prosodic features, Part-of-speech (POS), dialogue act (DA), repetitions, corrections, and syntactic-prosodic boundary to infer the emotion. For more details on such studies, readers are referred to the comprehensive survey of the past efforts in the field by Zeng & et al (2008). It must be noted, however, that although the above-mentioned studies reported an improved performance by using information of language, these systems typically depend on both accurate recognition of verbal content of emotional speech, which still cannot be reliably achieved by existing automatic speech recognition systems, and on accurate extraction of semantic discourse information, which is attained manually in the present systems.

6.4 Bio-potential features

Prior to extracting features, affect recognition systems that use bio-potential signals or modalities as input usually pre-process signals to remove noise (e.g., Savran & et al., 2006). Peripheral signals (e.g., GSR, respiration and blood volume pressure) are first filtered by a mean filter to remove noise (i.e., the resistance of the skin) or depending on the signal used, the environmental noise is removed by applying a bandpass filter. Various signal processing techniques such as Fourier transform, wavelet transform, thresholding, and peak detection, are commonly used to derive the relevant features from the physiological signals (Changchun & et al., 2005).

Following the preprocessing stage, there are various alternatives for feature extraction. For physiological signals, usually the following features are calculated: means, the standard deviations, the means of the absolute values of the first differences, the means of the absolute values of the first differences of the normalized signals, the means of the absolute values of the second differences, the means of the absolute values of the second differences of the normalized signals etc. (e.g., Picard & et al., 2001; Takahashi, 2004).

For brain signals, one alternative is to collect EEG energies at various frequency bands, time intervals and locations in the brain. The gathered signals are separated using frequency domain analysis algorithms and are then analyzed in terms of frequency bands (i.e., low, middle and high frequency band), and center frequency etc. (Takahashi, 2004). (Aftanas & et al., 2003, 2004) used the correlation between arousal variation and power in selected frequency bands and electrodes. Another possibility is to compute the STFT (Short Term Fourier Transform) on a pre-determined time segment of each electrode, assuming that the signal remains stationary within the chosen time widow (Savran & et al., 2006).

After these procedures, various pattern recognition techniques such as evaluation of subsets or feature selection, transformations of features, or combinations of these methods are applied. The extracted and calculated feature values then make up the overall feature vector used for classification.

Researchers reported that muscle activities (e.g., opening the mouth, clenching the jaw etc.) due to expression generation contaminate EEG signals with strong artifacts. Use of multiple sensors can thus cause cross-sensor noise (e.g., Savran & et al., 2006). Design of an appropriate filter or use of other techniques such as Independent component analysis (ICA) should be explored to remove this type of noise. Estimating a Laplacian reference signal by subtracting for each electrode the mean signal of its neighbors might potentially provide a better representation for the brain activity.

6.5 Thermal infrared imagery

Systems analyzing affective states from thermal infrared imagery perform feature extraction and selection, exploit temporal information (i.e., infrared video) and rely on statistical techniques (e.g., Support Vector Machines, Hidden Markov Models, Linear Discriminant Analysis, Principal Component Analysis etc.) just like their counterparts in visible spectrum imagery.

Current research in the thermal infrared imagery has utilized several different types of representations, from shape contours to blobs (Tsiamyrtzis & et al., 2007). Some studies estimate differential images between the averaged neutral face/body and the expressive face/body (e.g., Yoshitomi, 2000) and perform a transformation (e.g., discrete cosine transformation (DCT)). Other researchers divide each thermal image into grids of squares, and the highest temperature in each square is recorded for comparison (Khan & et al., 2006b). Patterns of thermal variations for each individual affective state are also used (Khan et al., 2006a). Similar tracking techniques to those in the visible spectrum are utilized (e.g., Condensation algorithm, Kalman/Particle Filtering etc.) and therefore, similar challenges pertain to tracking in the thermal infrared imagery domain (Tsiamyrtzis & et al., 2007).

7. Affective state recognition

The main problem overarching affect sensing is the recognition of affective states in their nature of complex spatio-temporal patterns. Should emotion recognition be regarded as an easier or a harder problem than an equivalent recognition problem in a *generic* domain? In the following, we identify its main characteristics as challenges or facilitators, alongside the main pattern recognition techniques that have been or can be used to deal with them.

7.1 Challenges

The main challenges we identify from the pattern recognition perspective can be listed as feature extraction, high inter-subject variation during recognition, dimensionality reduction, and optimal fusion of cues/modalities.

The value range of certain features is very limited compared to typical noise levels. Let us consider, for instance, the raising of an eyebrow that has to be recognized as an expression of surprise. When sensed by a camera, such a movement may translate into just a few pixels extent. Facial feature extraction from videos is typically affected by comparable errors, thus undermining recognition accuracy. Higher-resolution cameras (and lenses – in the order of several equivalent megapixels per frame) are required for effective feature extraction in such cases.

The space in which emotions have to be recognized is typically a feature space with very high dimensionality: for instance, (Gunes, 2007) uses a feature set with 152 face features and 170 upper-body features; (Bhatti2004) uses a feature set with 17 speech features; (Kim & et al, 2004) uses a feature set with 29 features from a combination of ECG, EMG, skin conductivity and respiration measurements. This aspect of emotion recognition is certainly a challenge and imposes the use of dimensionality reduction techniques. Linear combination techniques such as PCA and LDA and non-linear techniques such as KPCA have been used for that purpose, and so have been feature selection techniques such as Sequential Forward Selection (Bhatti & et al.,2004) and Genetic Algorithms (Noda & et al., 2006). However, feature-space dimensionality reduction for sequential data (not to be confused with reduction along the time dimension)is still an open problem.

High inter-subject variation is reported in many works. This challenges generalization over unseen subjects that are, most often, the actual targets of the emotion recognition process. The search for features with adequate discriminative power-invariance trade-off is an attempt at solving this problem (Varma & Ray, 2007).

Eventually, as modalities are heterogeneous and possibly asynchronous, their optimal fusion adds to the list of challenges. The asynchrony between modalities may be two fold: (a) asynchrony in subject's signal production (e.g., face movement might start earlier than the body movement), and (b) asynchrony during recording (e.g., a video recorded at 25 Hz frame rate while the audio recorded at 48 kHz) and processing of the signals coming from various sensing devices. For instance when fusing effective information coming from the EEG, the video or fNIRS, it should be noted that these have orders of magnitude difference in their relative time scales (Savran & et al., 2006).

The most straightforward approach to tackle modality fusion is at the decision or score level since feature- and time-dependence are abstracted. In decision level fusion each classifier processes its own data stream, and the two sets of outputs are combined at a later stage to produce the final hypothesis. There has been some work on combining classifiers and providing theoretical justification for using simple operators such as majority vote, sum, product, maximum/minimum/median and adaptation of weights (Kittler & et al., 1998). Decision-level fusion can also 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).

Recent works have attempted at providing synchronization between multiple cues to also support feature-level fusion, reporting greater overall accuracy when compared to decisionlevel fusion (e.g., Gunes, 2007 and Shan & et al., 2007). Feature-level fusion becomes more challenging as the number of features increases and they are of very different natures (e.g. distances and times). Synchronization then becomes of utmost importance.

Outside the affect sensing and recognition field, various techniques have been exploited for implicit synchronization purposes. For instance, dynamic time warping (DTW) has been used to find the optimal alignment between two time series if one time series may be warped non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between them. Variations of HMM have also been proposed for this task. The pair HMM model was proposed to align two non-synchronous training sequences and an asynchronous version of the Input/Output HMM was proposed for audio-visual speech recognition (Bengio, 2004). Coupled HMM and fused HMM have been used for integrating tightly coupled time series, such as audio and visual features of speech (Pan & et al., 2004). Bengio (Bengio, 2004), for instance, presents the Asynchronous HMM that could learn the joint probability of pairs of sequences of audiovisual speech data representing the same sequence of events (e.g., where sometimes lips start to move before any sound is heard for instance). 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 etc. (see Zheng & et al., 2008 for details on these).

A number of approaches have also been reported for explicit synchronization purposes. (Gunes, 2007) identifies the neutral-onset-apex-offset-neutral phases of face and body inputs and synchronizes the sequences at the phase level (i.e., apex phase). (Savran & et al., 2006) have obtained feature/decision level fusion of the fNIRS and EEG feature vectors and/or decision scores 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. Video features and fNIRS features can be fused at the feature or decision level on a block-by-block basis. (Paleari & Lisetti, 2006) introduce a generic framework with 'resynchronization buffers'. They aim to compare the different estimations, and realign the different evaluations so that they correspond to the same phenomenon even if one estimation is delayed compared to the other one.(e.g., in the case of delay).

For affective multimodal recognition, synchronization between the modalities is a very interesting and challenging problem and needs to be investigated further. In particular, synchronization other than feature-level, for example at higher levels of abstraction such as temporal phase or segment-level or even task-level synchronization, has not been be explored.

7.2 Facilitators

Affective data can be thought of as uninterrupted streams originating from a variety of sensors (cameras, microphones etc): prior to recognition, or simultaneously with it, it is also required to identify the data sequences corresponding to atomic emotions – a typical time segmentation problem in time series. In some applications, it is possible that a special neutral state can be recognized per se as the marker of the end of an emotion/start of the next, thus easing the time segmentation problem. This is the case, for instance, of a sequence of affective body gestures where each gesture concludes to an identifiable rest state.

Affective data are generated by humans under anatomical and biological constraints. This offers an unrivalled opportunity to simplify the recognition approach by exploiting such prior information. For instance, the generation of facial and bodily expressions undergo muscular constraints: a plateau is reached and maintained for a few seconds in which the features are at their maximum extent. (Gunes, 2007) uses this fact to decouple the temporal and spatial aspects of the recognition process: the plateau is identified first, prior to and independently of the specific emotion thanks to the constrained dynamics; emotion recognition is performed then by assuming that the feature values at the plateau are i.i.d. in the presence of noise. Similarly, (Elgammal & et al., 2003) posits a layer of "exemplars" that separate the spatial and temporal sides in a gesture recognition application. Use of such constraints should be incorporated in approaches to mitigate the high-dimensionality issue.

Affect is naturally expressed via multiple cues and channels. An adaptive framework based on fusion of the available cues and modalities thus offers an opportunity to improve the analysis and recognition of affective states. However, to date, most of the existing fusion algorithms have not been made adaptive to the input quality and therefore do not consider eventual changes on the reliability of the different information channels. (Paleari & Lisetti, 2006) proposed a generic fusion framework that is able to accept different single and multimodal recognition systems and to automatically adapt the fusion algorithm to find optimal solutions, and be adaptive to channel (and system) reliability. They describe a bufferized approach where two different fusion chains would be active in parallel. The first chain, treats close to real time signals and interpretations returning fast interpretations of the recognized emotion. The second chain works on the same bufferized and re-aligned data in order to have the possibility to resynchronize data just before fusion. The objective of this double chain is to have both a fast but less reliable and a longer but more accurate evaluations of the user's affective states.

Further research is needed to test the feasibility of the framework proposed by (Paleari & Lisetti, 2006) and/or create a more generic and common framework that can be easily adopted by the research community.

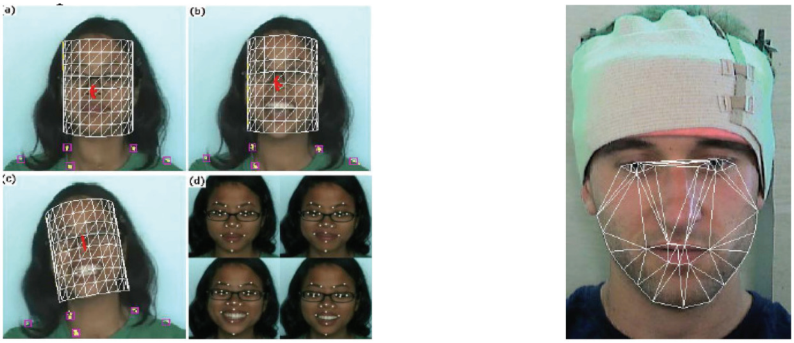
8. Affect sensing and recognition from multiple cues and modalities: representative systems

In this section we briefly review a number of automated systems that attempt to recognize affect from multicue or multimodal expressive behavior. This review is intended to be illustrative rather than exhaustive. For an exhaustive survey of past efforts in audiovisual affect sensing and recognition, the readers are referred to Zeng & et al. (Zeng & et al. , 2008).Here, we present representative projects/systems introduced in the multimodal affect recognition literature during the period 2004-2007 by grouping these systems under three categories: i) the lab, ii) from the lab to the real world and iii) the real world.

The lab systems analyze posed affect from multicue or multimodal expressive behavior. An example system is that of (Gunes, 2007) that recognizes 12 affective states (anger, anxiety, boredom, disgust, fear, happiness, negative surprise, positive surprise, neutral surprise, uncertainty, puzzlement and sadness) and their temporal segments (neutral-onset-apex-offset-neutral) from either face/upper-body/combined face-and-body display, acquired by two cameras simultaneously. The temporal segmentation of face and body display is achieved explicitly, a phase-synchronization scheme is introduced to deal with simultaneous yet asynchronous face and upper-body data and affective state recognition is performed both on a frame-basis and a video-basis. Experiments were conducted on the FABO database (Gunes & Piccardi, 2006a) from 10 subjects and 539 videos. The approach explores the usefulness of the temporal segment/phase detection to the overall task of affect recognition with various experiments. It also proposes fusion of information coming from multiple visual cues by phase synchronization and selective fusion, and proves the greater performance of this approach by comparative experiments. Using 50% of the data for training and remaining 50% for testing, the FABO system obtains an average recognition rate of 35% for facial expressions alone, 77% for bodily expression alone, %82.6 (frame-basis) and %85 (video-basis) by fusing face and upper-body data. From the experiments the authors concluded that explicit detection of the temporal phases can improve the accuracy of affective state recognition, recognition from fused face and body cues performs better than from facial or bodily expression alone, and synchronized feature-level fusion achieves better performance than decision-level fusion. (Shan & et al., 2007) also report affective state recognition results using the FABO database. They exploit the spatial-temporal features based on space-time interest point detection for representing body gestures in videos. They fuse facial expressions and body gestures at the feature-level by using the Canonical Correlation Analysis (CCA), a statistical tool that is suited for relating two sets of signals. For their experiments they selected 262 videos of seven affective states (anger, anxiety, boredom, disgust, happiness, puzzle, and surprise) from 23 subjects in the FABO database and obtained 88.5% recognition accuracy.

Systems that analyze (more) spontaneous or real world affect data from multiple cues or modalities are described as *from the lab to the real world systems*. An example system is that of (Valstar & et al., 2007). It automatically distinguishes between posed and spontaneous smiles by fusing information from multiple visual cues including the head, face, and shoulder actions. It uses a cylindrical head tracker to track the head motion; particle filtering with factorized likelihoods to track fiducial points on the face and auxiliary particle filtering to track the shoulders motion (see Figure 4a). Based on tracking data, the presence of AU6 (raised cheeks), AU12 (lip corners pulled up), AU13 (lip corners pulled up sharply), head movement (moved off the frontal view), and shoulder movement (moved off the relaxed state), are detected first. For each of these visual cues, the temporal segments (neutral, onset, apex, and offset) are also determined. Classification is then performed by combining

GentleBoost ensemble learning and Support Vector Machines (SVM). A separate SVM is trained for each temporal segment of each of the five behavioral cues (i.e., in total 15 GentleSVMs). The authors combined the results using a probabilistic decision function and investigated two aspects of multicue fusion: the level of abstraction (i.e., early, mid-level, and late fusion) and the fusion rule used (i.e., sum, product and weight criteria). Experimental results from 100 videos displaying posed smiles and 102 videos displaying spontaneous smiles were presented. Best results were obtained with late fusion of all cues when 94.0% of the videos were classified correctly (with 0.964 recall, and 0.933 precision). The results seem to indicate that using video data from face, head and shoulders increases the accuracy, and the head is the most reliable source, followed closely by the face.



(a) (b)

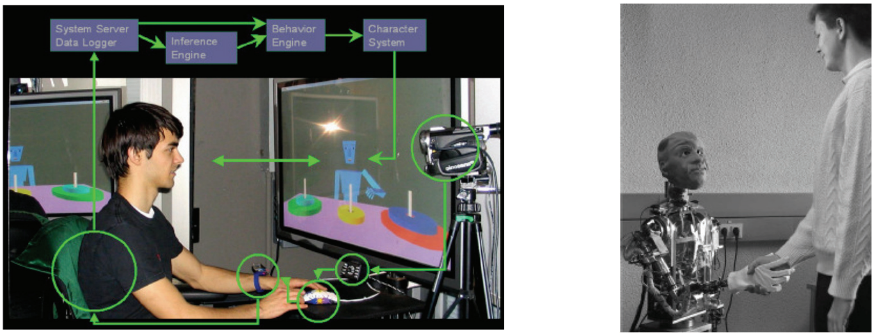
Figure 4. System of (a) (Valstar & et al., 2007) and (b) (Savran & et al., 2006).

(Savran & et al., 2006) present a project as part of the eNTERFACE Workshop on multimodal emotion detection from three modalities: brain signals via fNIRS, face video and the scalp EEG signals (see Figure 4b). fNIRS sensors were used to record frontal brain activity and EEG sensor was used to capture activity in the rest of the brain. In addition to these, a respiration belt, a GSR (Galvanic Skin Response) and a plethysmograph (blood volume pressure) were also used to record aperipheral body processes. All these devices were synchronized using a trigger mechanism. Three emotions (i.e., calm, exciting positive and exciting negative corresponding to neutral, happiness and disgust) were elicited in five subjects using emotionally evocative images evaluated on valence and arousal dimensions. Participants were then asked to self-assess their emotions by giving a score between 1 and 5 for valence and arousal components. For facial feature extraction an active contour-based technique and active appearance models (AAM) were used. For classification, Transferable Belief Model (TBM) was utilized. The authors considered fusion of fNIRS with video and of EEG with fNIRS. Fusion of all three modalities was not considered due to the extensive noise on the EEG signals caused by facial muscle movements. Both feature and decision level fusion was considered by adopting a block for each emotional stimuli (12.5 seconds long in their experiment) and a block-by-block fusion was applied. Assessment of emotion detection performance of individual modalities and their fusion has not been explored.

Takahashi proposed an emotion recognition system from multimodal bio-potential signals collected using an EEG sensor, a pulseoxymeter, and a skin conductance meter (Takahashi, 2004). Recordings of 12 subjects were obtained in a laboratory where the illumination, noise, and room temperature were controlled to maintain uniformity. To stimulate emotions (joy, anger, sadness, fear, and relax), several commercial films broadcasted on TV were used.

Recognition was carried out with NN and SVM using leave-one-out cross-validation method. The averaged recognition rate of 63.9% for NN and 66.7% for SVM was achieved. Pun & et al. describe the work they conducted in the domain of multimodal interaction via the use of EEG and other physiological signals for assessing a user's emotional status (Pun & et al., 2006). The experimental setup consisted of three participants viewing strongly positive or negative images. Ground truth consisted of the participant's self-assessment. The following physiological signals were recorded: EEGs, blood pressure, GSR, breathing rate, and skin temperature. Each participant was asked to provide valence and arousal values for each image they viewed. These values were then divided into either two classes (calm versus exciting for arousal, positive versus negative for valence), or three classes (same two classes plus an intermediate one). Features extracted such as signal power in particular frequency bands, means, standard deviations, and extreme values were saved as vectors. A Naïve Bayes classifier and a Fisher discriminant analysis were applied in a leave-one-out manner for classification. Depending on the classifier used, the participant, the use of either EEGs, or of peripheral signals only, or of both EEGs and peripheral signals, accuracies ranged between about chance level to 72% for the two classes problem, and between chance levels to 58% for the three classes problem.

Systems that analyze (more) realistic multimodal affect data are described as *the real world systems*. An example system is that of (Kapoor & et al., 2007) that assesses whether a learner is about to click on a button saying *I'm frustrated.* To this aim they use multiple nonverbal channels of information: a chair and a mouse both equipped with pressure sensors, a wireless skin conductance sensor placed on the wristband of the user, two cameras (one video camera for offline coding and the Blue-Eyes camera to record elements of facial expressions). See Figure 5a for details on the sensors used. The data obtained by the aforementioned sensors are classified into *pre-frustration* or *not pre-frustration* behavior using Gaussian process classification and Bayesian inference. The system deals with data synchronization in a similar manner to (Paleari & Lisetti, 2006). In other words, it gathers data for a predetermined time window (i.e., window size of 150 s), normalizes and then averages them. The proposed method was tested on data gathered from 24 participants using an automated learning companion. The experimental setup is described as follows. The users were asked to sit in front of a wide screen plasma display where an agent appears in a 3D environment. The user can interact with the agent and can attend to and manipulate objects and tasks in the environment. In the aforementioned experimental setup, the system was able to predict the indication of frustration from the collected data with 79% accuracy.



(a) (b)

Figure 5. (a) The system of (Kapoor & et al., 2007) and (b) a humanoid interacting in a humanlike manner (Spexard & et al., 2007).

(Spexard & et al., 2007)present an anthropomorphic robot framework (BARTHOC) bringing together different interaction concepts and perception capabilities with the goal of creating an interdisciplinary research platform for multimodal human-robot interaction (HRI). The framework uses two cameras and two microphones only, has components for face detection, a tracking module based on anchoring, and extended interaction capabilities based on both verbal and nonverbal communication (see Figure 5b). Sounds are validated as voices by using the results of a face detector. The robot is equipped with an attention system for tracking and interacting with multiple persons simultaneously in real time. As sensors cover a limited area only, people are tracked by utilizing a short-time person memory that extends the existing anchoring of people. A long-time memory stores person specific data into file enabling robust tracking in real time. A modular integration approach utilizing XML-based data exchange is used for implementing different interaction capabilities like deictic gestures, natural adaptive dialogs, and emotion awareness on the robot. The robot can recognize affect by classifying the prosody of an utterance to seven emotional states (happiness, anger, fear, sadness, surprise, disgust, and boredom) independently from the content in emotional states of the speaker. The robot is thus able to realize when a communication partner is getting angry and can react showing a calming facial expression on its face. The appropriate facial expression can be invoked from different modules of the overall system, e.g., BARTHOC starts smiling when it is greeted by a human and "stares" onto an object presented to it. The framework also contains a 3-D body tracker based on 2-D video data and a 3-D body model to compensate the missing depth information from the video data. Deictic gestures and the position a person is referring to are estimated using the direction and speed of the body extremity trajectories. The robot is then able to perform pointing gestures to presented objects itself. Robot's emotion recognition and facial expression generation capabilities were evaluated by creating a setup in which multiple persons were invited to read out a shortened version of the fairy tale to the robot. For this experiment, an office-like surrounding with common lighting conditions was used. The robot mirrored the classified prosody of the utterances during the reading in emotion mimicry at the end of any sentence, grouped into happiness, fear, and neutrality. As the neutral expression was also the base expression, a short head movement toward the reader was generated as a feedback for non-emotional classified utterances. Overall, the use of emotion recognition and mimicry of the robot was found to be encouraging for further research in a robotic platform for multimodal human-robot interaction.

9. Conclusion and discussion

This chapter focused on the challenges faced when moving from affect recognition systems that were designed and experimented in laboratory settings (i.e., analyzing posed data) to the real world systems (i.e., analyzing spontaneous data) in a multicue and/or multimodal framework. It discussed the problem domain of affect sensing and recognition by explicitly focusing on multiple input modalities (audio, vision, tactile, and thought) and cues (facial expressions, head and body gestures, etc.) together with alternative channels (brain and thermal infrared signals), and explored a number of representative systems introduced during the period 2004-2007, either capable of handling laboratory, more realistic or real world settings.

The analysis provided in this chapter indicates that the automatic multimodal affect recognition systems have slowly but steadily started shifting their focus from the lab to the real world settings. There already exist a number of efforts for automatic multimodal affect recognition in real world settings. Existing systems deal with the so-called spontaneous data obtained in less-controlled or restricted environment (i.e., subjects are taking part in the interaction, subjects are not always stationary, etc.), and can handle a limited number of emotion categories (e.g., 2-6). These real world systems have been mostly trained to have expertise in a specific interaction context. As stated by (Kapoor & et al., 2007), generalization thus might be affected by various factors such as: the experimental setup (i.e., the tasks and situations the users are presented), age of the users, availability/robustness of the sensors (e.g. the skin conductance sensor is effected by sweat) etc.

One of the main disadvantages of the bio-potential 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 the simultaneous use of these modalities needs to be reconsidered. Additionally, during recording the fNIRS device is known to cover the eyebrows. This in turn poses another challenge (i.e., occluding facial features) for multimodal affective data recordings if the simultaneous use of these modalities is intended. However, new forms of non-contact psychological sensing might help spreading the use of psychological signals as input to multimodal affect recognition systems.

The most notable issue is the fact that there exists a gap between different communities researching emotions or affective states. For instance, affect recognition communities seem to use different databases compared to psychology or cognitive science communities. Moreover, for annotation of the data, a more uniform and multi-purpose scheme that can accommodate all possible modalities should be explored. Another issue to consider is fusion of multimodal affect data. Researchers claim that the choice of fusion strategy depends on the targeted application (Wu & et al., 1999, Busso & et al., 2004). Accordingly, all available multimodal recognizers have designed and/or used ad hoc solutions for fusing information coming from multiple modalities but cannot accept new modalities. In summary, there is not a general consensus when fusing multiple modalities.

An important point to note is that experimentation with all possible human behavioral cues (linguistic terms/words, audio cues such as pitch, facial expression/AUs, body postures and gestures, physiological signals, brain and thermal infrared signals etc.) has been impossible to date due to lack of a generic and shared platform for automatic affect recognition. We would like to stress that it is highly likely that machines aimed at assisting human users in their tasks will need neither the human-like flexibility to adapt to any environment and any situation nor will they need to function exactly as humans do. Machines can be highly profiled for a specific purpose, scenario, user, etc. Nonetheless, it is necessary to investigate which modalities are the most suitable for which application context. The representative systems covered in this chapter are thus encouraging towards such a goal.

However, further research is still needed in order to identify the importance/feasibility of the following questions/factors for creating multimodal affect recognizers that can handle the so-called more natural or real world settings:

- Among the available *external* and *internal* modalities, which ones should be used for automatic affect recognition? In which context? Will the affect recognition accuracy increase as the number of modalities a system can analyze/integrate increases?
- In automated multimodal affect systems, can global processing replace local processing (i.e., whole-body expression analysis instead of facial expression analysis) by still providing means for fast and accurate analysis (e.g., when distance between the subject and the cameras/sensors poses a challenge)?
- What cross-modal interactions between pairs of various modalities (e.g., tactile and visual, tactile and audio etc.) can be exploited for multimodal affect analysis? Can we follow the example of HHI where judgments for one modality are influenced by a second modality even at the cost of increased ambiguity? How can such analysis be integrated for fusion of modalities?
- How can automated systems detect and label an affective message conveyed by different modalities as either *congruent* (i.e., agreeing) or *incongruent* (i.e., disagreeing)? After labeling, how can such knowledge be incorporated into the multimodal systems for detailed understanding of the information being conveyed? Should the goal be towards detecting and decreasing ambiguity, and increasing the reliability and accuracy of the automatic recognition process? Should/can we use the so-called *internal signals* (e.g., thermal infrared or physiological signals) for resolving ambiguity, instead of relying purely on the *external ones*?
- For the fusion purposes, how can an automated system include and integrate a new modality (when it becomes available) automatically? How can the system dynamically adapt to the channel conditions (e.g., when noise increases) in order to find an optimal solution?
- For the recognition purposes, how can a system estimate different affective phenomena (emotions, moods, affects and/or personalities)? How should the system include the knowledge about the environment and the user to the overall multimodal recognizer?
- How should the requirements of an automated system be decided? Is real-time processing and outputting labels as quickly as possible the priority? Or is the priority having a better, more accurate understanding of the user’s affective state, regardless of the computational time it will take (Paleari & Lisetti, 2006)?

Overall, the research field of multimodal affect sensing and recognition is relatively new, and future efforts have to follow to address the aforementioned questions.

10. Acknowledgement

The authors would like to thank Arman Savran, Bulent Sankur, Guillaume Chanel, Rana el Kaliouby, Rosalind Picard and Thorsten Spexard for granting permission to use figures from their works/papers. The research of Maja Pantic leading to these results has been funded in part by the European IST Programme Project FP6-0027787 (AMIDA) and the European Community's 7th Framework Programme [FP7/2007-2013] under grant agreement no 211486 (SEMAINE).

11. References

Affect analysis group (2008): http://www.pitt.edu/%7Eemotion/publications.html.

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*, Vol. 50, No. 3, 205-212.

Aftanas, L.I.; Reva, A.A.; Varlamov; 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*, (859-867).

Ali, A. N. & Marsden, P. H. (2003). Affective multi-modal interfaces: the case of McGurk effect, *Proc. of the 8th Int. Conf. on Intelligent User Interfaces*, pp. 224 - 226.

Allwood, J. & et al. (2004), The MUMIN multimodal coding scheme, *Proc. Workshop on Multimodal Corpora and Annotation*.

Ambady, N. & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, Vol. 64, 431-441.

Ambady, N. & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta–analysis. *Psychological Bulletin,* Vol. 11, No. 2, 256–274.

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

Ashraf, A.B.; Lucey, S.; Cohn, J.F.; Chen, T.; Ambadar, Z.; Prkachin, K.; Solomon, P.; & Theobald, B. J. (2007). The painful face: Pain expression recognition using active appearance models. *Proc. of the ACM Int. Conf. on Multimodal Interfaces*, pp. 9-14.

Banziger, T. & Scherer, K. R. (2007) Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus, *Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction*, pp. 476–487.

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

Batliner, A.; Fischer K.; Hubera, R.; Spilkera, J. & Noth, E. (2003). How to find trouble in communication. *Speech Communication*, Vol. 40, 117-143.

Bengio, S. (2004). Multimodal speech processing using asynchronous hidden markov models, *Information Fusion*, Vol. 5, 81–89.

Bhatti, M.W.; Yongjin Wang & Ling Guan (2004). A neural network approach for human emotion recognition in speech, *Proc. International Symposium on Circuits and Systems*, Vol. 2, pp. 181-184.

Buller, D.; Burgoon, J.; White, C. & Ebesu, A. (1994). Interpersonal deception: Vii. behavioral profiles of falsification, equivocation and concealment. *Journal of Language and Social Psychology*, Vol. 13, No. 5, 366–395.

Busso, C.; Deng, Z.; Yildirim, S.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Lee, S.; Neumann, U. & Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information, *Proc. Int. Conf. on Multimodal Interfaces*, pp. 205–211.

Cacioppo, J. T. & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates, *Psychological Bulletin,*Vol. 115, 401-423.

Campbell, N. & Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension, *Proc. Int'l Congress of Phonetic Sciences*, pp. 2417-2420.

Camras, L.A.; Meng, Z. ; Ujiie, T. ; Dharamsi, K.; Miyake, S.; Oster, H.; Wang, L.; Cruz, A.; Murdoch, J. & Campos, J. (2002). Observing emotion in infants: facial expression, body behavior, and rater judgments of responses to an expectancy-violating event, *Emotion*, Vol. 2, 179-193.

Camurri, A.; Mazzarino, B. & Volpe, G. (2003) Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library. *Proc. of Gesture Workshop*, pp. 460-467.

Cassell, J. (1998). A framework for gesture generation and interpretation, In: *Computer Vision in Human-Machine Interaction*, A. Pentland & R. Cipolla (Ed.), Cambridge University Press.

Changchun Liu; Rani, P. & Sarkar, N. (2005). An empirical study of machine learning techniques for affect recognition in human-robot interaction, *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2662- 2667.

Coulson, M. (2004). Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence, *Nonverbal Behavior*, Vol. 28, No. 2, 117–139.

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W. & Taylor, J.G. (2001). Emotion Recognition in Human-Computer Interaction, *IEEE Signal Processing Magazine*, Vol. 18, No. 1, 32-80.

Darwin, C. (1872). *The expression of the emotions in man and animals*, John Murray, London.

De Gelder, B.; Bocker, K. B.; Tuomainen, J.; Hensen, M. & Vroomen, J. (1999). The combined perception of emotion from voice and face: Early interaction revealed by human electric brain responses, *Neuroscience Letters*, Vol. 260, 133–136.

DePaulo, B. Cues to deception (2003). *Psychological Bulletin*, Vol. 129, No. 1, 74–118.

Devillers, L.; Vasilescu, I. & Vidrascu, L. (2004). Anger versus fear detection in recorded conversations, *Proc. Speech Prosody*, pp. 205-208.

Douglas-Cowie, E.; Cowie, R., Sneddon; Cox, C.; Lowry, McRorie, Martin, J.-C.; Devillers, L. & Batliner, A. (2007). The HUMAINE Database: addressing the needs of the affective computing community, *Proc. of the Second International Conference on Affective Computing and Intelligent Interaction*, pp. 488-500.

Dreuw, P.; Deselaers, T.; Rybach, D.; Keysers, D. & Ney, H. Tracking using dynamic programming for appearance-based sign language recognition (2006). *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 293–298.

Driver, J. & Spence, C. (2000). Multisensory perception: Beyond modularity and convergence, *Current Biology*, Vol. 10, No. 20, 731-735.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of New York Ac. of sciences*, Vol. 1000, 105–221.

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

Ekman, P.(1979) About brows: Emotional and conversational signals, In: *Human Ethology: Claims and Limits of a New Discipline: Contributions to the Colloquium*, M.V. Cranach, K. Foppa, W. Lepenies, and D. Ploog (Ed.), 169–248, Cambridge University Press, New York.

Ekman, P. & Friesen, W.V. (2003) *Unmasking the face: a guide to recognizing emotions from facial clues*. Cambridge, MA.

Ekman, P. & Friesen, W. V. (1978). *Facial action coding system: A technique for the measurement of facial movement*, Palo Alto, Calif.: Consulting Psychologists Press.

Elgammal, A.; Shet, V.; Yacoob, Y. & Davis, L.S. (2003). Learning dynamics for exemplar-based gesture recognition, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 571–578.

El Kaliouby, R. & Teeters, A. (2007) Eliciting, capturing and tagging spontaneous facial affect in autism spectrum disorder, *Proc. of the 9th Int. Conf. on Multimodal Interfaces*, pp. 46-53.

El Kaliouby, R. & Robinson, P. (2005). Real-time Inference of Complex Mental States from Facial Expressions and Head Gestures, *In: Real-Time Vision for HCI*, pp. 181–200, Spring-Verlag.

Fasel, I. R.; Fortenberry, B. & Movellan, J. R. (2005). A generative framework for real-time object detection, and classification, *Computer Vision and Image Understanding*, Vol. 98.

Fragopanagos, F. & Taylor, J.G., Emotion recognition in human-computer interaction, *Neural Networks*, Vol. 18, 389-405.

Friesen, W. V., & Ekman, P. (1984). EMFACS-7: *Emotional Facial Action Coding System*, Unpublished manuscript, University of California at San Francisco.

Gross, M. M.; Gerstner, G. E.; Koditschek, D. E.; Fredrickson, B. L. & Crane, E. A. (2006) Emotion Recognition from Body Movement Kinematics:http://sitemaker.umich.edu/mgrosslab/files/abstract.pdf.

Gunes, H. (2007) Vision-based multimodal analysis of affective face and upper-body behaviour, Ph.D. dissertation, University of Technology, Sydney (UTS), Sydney, Australia.

Gunes, H. & Piccardi, M. (2008). From Mono-modal to Multi-modal: Affect Recognition Using Visual Modalities, In: *Ambient Intelligence Techniques and Applications*, D. Monekosso, P. Remagnino, and Y. Kuno (Eds.), Springer-Verlag (in press).

Gunes, H. & Piccardi, M. (2007). Bi-modal emotion recognition from expressive face and body gestures. *J. Network and Computer Applications,* Vol. 30, No. 4, 1334-1345.

Gunes, H. & Piccardi, M. (2006a).A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior, *Proc. of the Int. Conf. on Pattern Recognition*, Vol. 1, pp. 1148–1153.

Gunes, H. & Piccardi, M. (2006b), Creating and annotating affect databases from face and body display: A contemporary survey, *Proc. of the IEEE Int. Conf. on Systems, Man and Cybernetics*, pp. 2426–2433.

Hadjikhani, N. & De Gelder, B. (2003). Seeing fearful body expressions activates the fusiform cortex and amygdala, *Current Biology*, Vol. 13, 2201-2205.

Jenkins, J.M.; Oatley, K. & Stein, N.L. (1998). *Human emotions: A reader*, Blackwell Publishers, Malden, MA.

Juslin, P.N. & Scherer, K.R. (2005), Vocal expression of affect, In: *The New Handbook of Methods in Nonverbal Behavior Research*, Harrigan, J., Rosenthal, R. & Scherer, K. (Ed.), Oxford University Press, Oxford, UK.

Kapoor, A.; Burleson, W. & Picard, R. W. (2007). Automatic Prediction of Frustration, *Int. Journal of Human Computer Studies,* Vol. 65, No. 8, 724-736.

Karpouzis, K.; Caridakis, G.; Kessous, L.; Amir, N.; Raouzaiou, A. ; Malatesta, L. & Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal and bodily expressions recognition, In: *Lecture Notes in Artificial Intelligence*, vol. 4451, pp. 92–116.

Khan, M. M.; Ingleby, M. & Ward, R. D. (2006a). Automated facial expression classification and affect interpretation using infrared measurement of facial skin temperature variations, *ACM Transactions on Autonomous and Adaptive Systems*, Vol. 1, No. 1, 91 – 113.

Khan, M. M.; Ward, R. D. & Ingleby, M. (2006b). Infrared Thermal Sensing of Positive and Negative Affective States, Proc. of the IEEE Conf. on Robotics, Automation and Mechatronics, pp. 1-6.

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

Kittler, J.; M.; Hatef, Duin, R.P.W. & Matas, J. (1998). On combining classifiers, *IEEE Tran. on Pattern Analysis and Machine Intelligence,* Vol. 20, No. 3, 226–239.

Laban, R. & Ullmann, L. (1988). *The mastery of movement*, 4th revision ed., Princeton Book Company Publishers.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention, *American Psychologist*, Vol. 50, No. 5, 372–385.

Lienhart, R. & Maydt, J. (2002). An extended set of haar–like features for rapid object detection, *Proc. of the IEEE Int. Conf. on Image Processing*, Vol. 1, pp. 900–903.

Littlewort, G. C.; Bartlett, M. S. & Lee, K. (2007). Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain, *Proc. of the 9th Int. Conf. on Multimodal interfaces*, pp. 15-21.

Martin, J. -C.; Caridakis, G.; Devillers, L.; Karpouzis, K. & Abrilian, S. (2007). Manual annotation and automatic image processing of multimodal emotional behaviors: validating the annotation of TV interviews, *Personal and Ubiquitous Computing*.

Martin, J. -C.; Abrilian, S. & Devillers, L. (2005). Annotating Multimodal Behaviours Occurring During Non Basic Emotions, *Proc. of the First Int. Conf. on Affective Computing and Intelligent Interaction*, pp. 550–557.

Massaro, D. W. & Cohen, M. M. (2000), Fuzzy logical model of bimodal emotion perception: Comment on "The perception of emotions by ear and by eye" by de Gelder and Vroomen, *Cognition and Emotion*, Vol. 14, No. 3, pp. 313-320.

McNeill, D. (1985). So you think gestures are nonverbal?, *Psychological Review*, Vol. 92, 350-371.

Meeren, H. K.; Van Heijnsbergen, C. C. & De Gelder, B. (2005). Rapid perceptual integration of facial expression and emotional body language, *Proc. of the National Academy of Sciences of the USA*, Vol. 102, 16518–16523.

Mitra, S. & Acharya, T. (2007). Gesture Recognition: A Survey, *IEEE Tran. on Systems, Man, and Cybernetics, Part C*, Vol. 37, No. 3, 311-324.

Nakasone, A.; Prendinger, H. & Ishizuka, M. (2005). Emotion Recognition from Electromyography and Skin Conductance, *Proc. of the 5th International Workshop on Biosignal Interpretation*, Tokyo, Japan, 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*, Vol. 84, 783-790.

Ning, H.; Han, T.X.; Hu, Y.; Zhang, Z.; Fu, Y. & Huang, T.S. (2006). A real-time shrug detector, *Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition*, pp. 505–510.

Noda, T.; Yano, Y.; Doki, S. & Okuma, S. (2006). Adaptive Emotion Recognition in Speech by Feature Selection Based on KL-divergence, *Proc. IEEE International Conference on Systems, Man and Cybernetics*, Vol. 3, pp. 1921 - 1926.

Ortony, A. & Turner, T. J. (1990). What's basic about basic emotions?, *Psychological Review*, Vol. 97, pp. 315–331.

O'Toole, A. J. & et al. (2005). A video database of moving faces and people, *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 27, No. 5, 812-816.

Paleari, M. & Lisetti, C. L. (2006). Toward multimodal fusion of affective cues, *Proc. of the 1st ACM Int. Workshop on Human-Centered Multimedia*, pp. 99–108.

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*, Vol. 52, No. 3, 573–581.

Pantic, M. & Bartlett, M.S. (2007). Machine Analysis of Facial Expressions, In: *Face Recognition*, Delac, K. & Grgic, M. (Ed.), Vienna, Austria: I-Tech Education and Publishing, 377-416.

Pantic, M.; Pentland, A.; Nijholt, A. & Huang, T. (2007). Machine understanding of human behavior, In: *Lecture Note in Artificial Intelligence*, Vol. 4451, pp. 47-71.

Pantic, M. & Rothkrantz, L.J.M. (2003). Towards an Affect-Sensitive Multimodal Human-Computer Interaction, *Proceedings of the IEEE*, Vol. 91, No. 9, 1370-1390.

Pavlidis, I. T.; Levine, J.; Baukol, P. (2001). Thermal image analysis for anxiety detection, *Proc. of the International Conference on Image Processing*, Vol. 2, pp. 315 - 318.

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

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*, Vol. 23, No. 10, 1175-1191.

Poppe, R. (2007). Vision-based human motion analysis: An overview, *Computer Vision and Image Understanding*, Vol. 108, No. 1-2, 4-18.

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 Tran. on Neural Systems and Rehabilitation Engineering*, Vol. 14, No. 2.

Puri, C.; Olson, L.; Pavlidis, I.; Levine, J. & Starren, J. (2005). StressCam: Non-contact measurement of users' emotional states through thermal imaging, Proc. of the CHI, pp. 1725-1728.

Riseberg, J.; Klein, J.; Fernandez, R. & Picard, R.W. (1998). Frustrating the User on Purpose: Using Biosignals in a Pilot Study to Detect the User's Emotional State, *Proc. of CHI.*

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

Russell, J.A. & Carroll, J. M. (1999). On the bipolarity of negative and positive affect, Psychological Bulletin 125 (1999), 3-30.

Russell, J.A. & Fernández-Dols, J.M., (1997). The Psychology of Facial Expression. New York: Cambridge Univ.

Savran, A.; Ciftci, K.; Chanel, G.; Mota, J. C.; Viet, L. H.; Sankur, B.; Akarun, L.; Caplier, A. & Rombaut, M. (2006). Emotion Detection in the Loop from Brain Signals and Facial Images, *Proc. of the eNTERFACE 2006*, July 17th – August 11th, Dubrovnik, Croatia, Final Project Report (www.enterface.net).

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

Shan, C.; Gong, S. & McOwan, P. W. (2007). Beyond facial expressions: Learning human emotion from body gestures, *Proc. of the British Machine Vision Conference.*

Spexard, T. P.; Hanheide, M. & Sagerer, G. (2007). Human-Oriented Interaction With an Anthropomorphic Robot, *IEEE Tran. on Robotics*, Vol. 23, No. 5.

Takahashi, K. (2004). Remarks on Emotion Recognition From Multi-Modal Bio-Potential Signals, *Proc. IEEE International Conference on Industrial Technology*, pp. 1138-1143.

Tian, Y. L.; Kanade, T. & Cohn, J. F. (2002), Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity, *Proc. of the IEEE Int. Conf. on Automaitc Face and Gesture Recognition*, pp. 218-223.

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,* Vol. 71, No. 2, 197-214.

Valstar, M. F.; Gunes, H. & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features, *Proc. of the 9th Int. Conf. on Multimodal Interfaces*, pp. 38–45.

Van den Stock J.; Righart R. & De Gelder B. (2007). Body expressions influence recognition of emotions in the face and voice, *Emotion*, Vol. 7, No. 3, 487-494.

Van Hoof, J.A. (1962). Facial expressions in higher primates, *Proceedings of the Symposium of the Zoological Society of London*, Vol. 8, pp. 97-125.

Varma, M. & Ray, D. (2007). Learning The Discriminative Power-Invariance Trade-Off, *Proc. IEEE 11th International Conference on Computer Vision*, pp. 1-8.

Vianna, D.M. & Carrive, P. (2005). Changes in cutaneous and body temperature during and after conditioned fear to context in the rat, *European Journal of Neuroscince*, Vol. 21, No. 9, 2505-25012.

Villalba, S. D. ; Castellano, G. & Camurri, A. (2007). Recognising human emotions from body movement and gesture dynamics, *Proc. of the Second Int. Conf. on Affective Computing and Intelligent Interaction*, pp. 71–82.

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518.

Vroomen, J.; Driver, J. & De Gelder, B. (2001). Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive, Affective & Behavioral Neuroscience*, Vol. 1, 382–387.

Walker-Andrews, A. S. (1997). Infants' perception of expressive behaviors: differentiation of multimodal information, *Psychol Bull*, Vol. 121, No. 3, 437-456.

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

Wilson, A. D.; Bobick, A. F. & Cassell, J. (1997). Temporal classification of natural gesture and application to video coding, *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 948–954.

Wu, L.; Oviatt, S.L. & Cohen, P.R. (1999). Multimodal integration–a statistical view, *IEEE Tran. on Multimedia,* Vol. 1, No. 4, 334–341.

Yilmaz, A.; Javed, O. & Shah, M. (2006). Object Tracking: A Survey, *ACM Journal of Computing Surveys*, Vol. 38, No. 4.

Yoshitomi, Y.; Kim, S.-I.; Kawano, T. & Kilazoe, T. (2000) Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, *Proc. of the IEEE International Workshop on Robot and Human Interactive Communication*, pp. 178 – 18.

Zeng, Z.; Pantic, M.; Roisman, G.I. & Huang, T.S. (2008). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, *IEEE Tran. on Pattern Analysis and Machine Intelligence*, 2008 (in press).

Zuckerman, M.; Larrance, D. T.; Hall, J. A.; DeFrank, R. S. & Rosenthal, R. (1979). Posed and spontaneous communication of emotion via facial and vocal cues, *Journal of Personality*, Vol. 47, No. 4, 712–733.