Categorical and dimensional affect analysis in continuous input: Current trends and future directions☆

Hatice Gunes [a,*], Björn Schuller [b]

[a] School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[b] Institute for Human-Machine Communication, Technische Universität München, Germany

1. Introduction: continuous input

Continuity is in the core of everything that is alive. Consciously or not, aware or not, human beings breath continuously, think and (re) act continuously. In an era where human and computer existence has become extremely interwoven, there is a need to somehow incorporate the continuity aspect into Human–Computer Interaction (HCI). Recent developments in hardware (e.g., depth cameras) and software (e.g., computer vision and scene analysis) have already paved the way toward HCI settings that are based on touch, gesture, and movement (e.g., Microsoft Kinect) as well as advanced direct manipulation. Means and modes of input also slowly but steadily move away from discrete actions toward continuous input.

In the context of affective human behavior analysis, we use the term continuous input to refer to naturalistic settings where explicit or implicit input from the subject is continuously available. In other words, in an interaction setting a human subject at times may play the role of a producer of the communicative behavior, while at other times the role of a recipient of the communicative behavior, with appropriate turn-taking behavior. As a result, the analysis and the response provided by the automatic system are also envisioned to be continuous over the course of time, within the boundaries of digital machine output (see the illustrations in Figs. 1 and 2) [3]. This does not necessarily mean that the user is always actively engaged with the system. Instead, it refers to settings where the system is alive and available via continuous user presence and nonverbal signal analysis, and appropriate responsiveness.

The term continuous affect analysis is used as analysis that is continuous in time as well as analysis that uses affect phenomenon represented in dimensional space. The former refers to acquiring and processing long unsegmented recordings for detection of an affective state or event (e.g., nod, laughter, pain), and the latter to the prediction of an affect dimension (e.g., valence, arousal, power).

Needless to say, continuity in input, analysis and synthesis brings along a number of challenges. The first challenge is the requirement for soft real-time processing and responsiveness. Automatic systems are expected to be responsive to continuous input provided by the user in real-time (e.g., an immediate visual or audio response). The expectation holds even if full processing and analysis of the input takes somewhat longer time to compute. Therefore the current trend is to make an impression on the user that the application responds well and in a timely manner, and provides an almost immediate and in-place feedback (soft real-time rather than full real-time [4–7]).

In light of these arguments, this special issue focuses on affect analysis in continuous input and aims at discussing the issues and

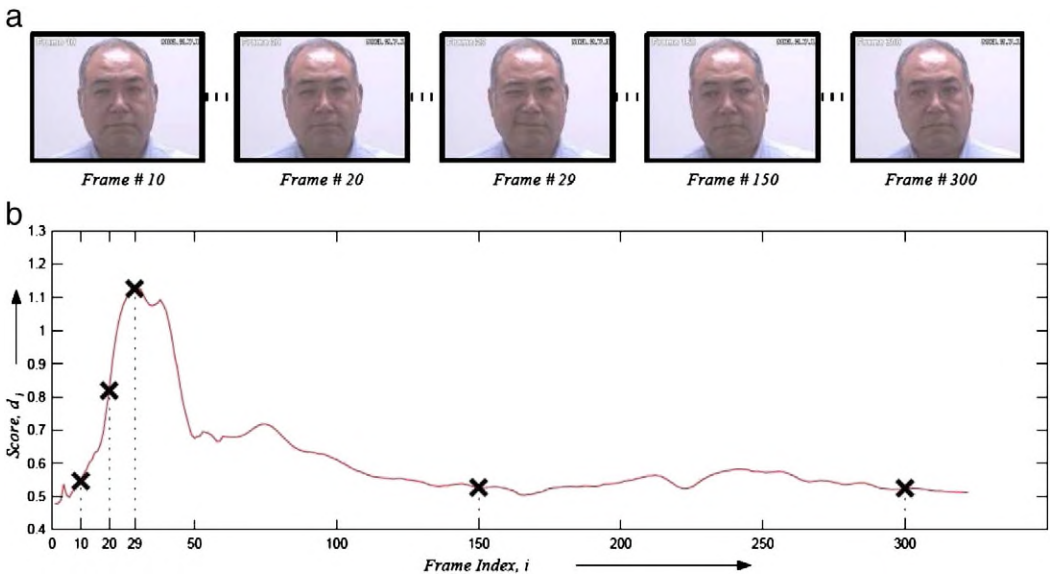

Fig. 1. A representative example of continuous affect detection, in this case detection of pain [1,2]. (a) Sample frames from a pain-video sequence with their frame indices, (b) scores for individual frames. Points corresponding to the frames shown in (a) are highlighted. For the example sequence shown, a cumulative score was computed and compared to the derived equal error threshold of −3 to yield an output decision of *pain*.
The figure is courtesy of [1].

the challenges pertinent in sensing, recognizing and responding to continuous human affective behavior from diverse communicative cues and modalities.[1] In line with the special issue, this survey paper aims to put the continuity aspect of affect under the spotlight by investigating the current trends in affect analysis in continuous input, and provides guidance towards possible future directions. The paper differs from the previous survey papers that investigated dimensional and continuous affect analysis and synthesis (e.g., [8–10]) by focusing on the latest developments and trends, and by incorporating the works that focus on continuous input and analysis regardless of the discrete or continuous emotion representation model used.

The paper is structured as follows: we first focus on discrete and continuous affect representation approaches (Section 2). We provide a generalized view of affect analysis in continuous input in Section 3. We then proceed with exploring the problem domain of affect analysis in continuous input by focusing on communicative modalities and cues (Section 4), data acquisition and annotation (Section 5), and automatic analysis and prediction (Section 6). Section 7 presents available frameworks and tools that can potentially be used for affect analysis in continuous input. Section 8 focuses on recent applications introduced in relevant fields. Section 9 provides a representative description of recently organized competitions, and journal Special Issues and book compilations published. The paper concludes by discussing the future trends and providing some recommendations to advance the field (Section 10).

2. Affect representation

How to represent emotions and affect is one of the first decisions to be made prior to creating an automatic affect analyzer. We provide details and discussion on affect representation approaches by grouping them under two categories: (i) social sciences approach and (ii) social sciences approach applied in engineering.

2.1. The social sciences approach

There exist various theories on how to represent affect. The most widely accepted approaches for modeling affect are the categorical, the dimensional, and the appraisal-based approach [11]. The categorical approach claims that there exist a small number of emotions that are

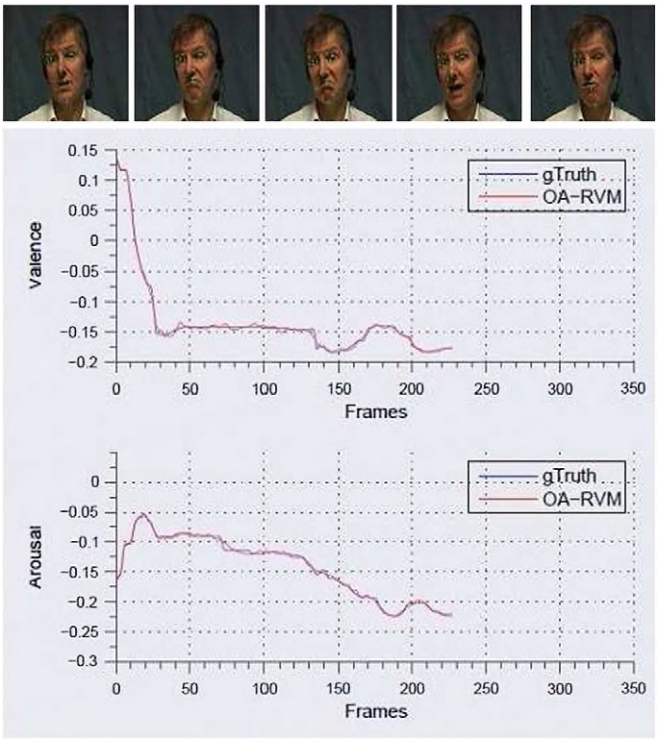

Fig. 2. Illustration of how continuous prediction of valence and arousal dimensions is achieved from tracked facial feature points. In this case, visual data is incoming continuously, facial feature points are tracked and dimensional affect prediction is provided for each and every frame in a continuous scale via Output-Associative Relevance Vector Machine (OA-RVM). gTruth: ground truth annotation, OA-RVM: prediction provided by the OA-RVM regressor.

[1] We use the distinction provided by the Oxford dictionary that defines modality as a particular form of sensory perception (e.g., the visual and auditory modalities), and cue as a feature of something perceived that is used in the brain's interpretation of the perception (e.g., expectancy is communicated both by auditory and visual cues).

basic, hard-wired in our brain, and recognized universally (e.g., [12]). This theory on universality and interpretation of affective nonverbal expressions in terms of basic emotion categories has been the most commonly adopted approach in research on automatic measurement of human affect. However, a number of researchers have shown that in everyday interactions people exhibit non-basic, subtle and rather complex affective states like thinking, embarrassment or depression. 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 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 [13]. These states were found relevant in representation for affect detection (e.g., [14]). Such subtle and complex affective states can be expressed via dozens of anatomically possible facial and bodily expressions, audio or physiological signals. Therefore, a single label or any small number of discrete classes may not reflect the complexity of the affective state conveyed by such rich sources of information [15]. Hence, a number of researchers advocate the use of dimensional description of human affect, where affective states are not independent from one another; rather, they are related to one another in a systematic manner (e.g., [11,16,15,17]). In the categorical approach, where each affective display is classified into a single category, complex mental/affective states or blended emotions may be too difficult to handle [18]. Instead, in the dimensional approach, emotion transitions can be easily captured, and observers can indicate their impression of moderate and naturalistic emotional expressions on several continuous scales. Hence, dimensional modeling of emotions has proven to be useful in several domains (e.g., affective content analysis [19], creating sensitive artificial listeners [5]).

The most widely used dimensional model is a circular configuration called Circumplex of Affect [15]. This model is based on the hypothesis that each basic emotion represents a bipolar entity being a part of the same emotional continuum. The proposed polars are arousal (relaxed vs. aroused) and valence (pleasant vs. unpleasant). Using a dimensional representation has a number of advantages over the categorical representation. The circumplex model allows the representation of emotion intensity, as well as similarity and contrast between various emotion categories. Another well-accepted and commonly used dimensional description is the 3D-emotional space of pleasure–displeasure, arousal–non-arousal, and dominance–submissiveness [16], at times referred to as the PAD emotion space [20] or as emotional primitives [21]. To guarantee a more complete description of affective coloring, some researchers include expectation as the fourth dimension [22], and intensity as the fifth dimension (e.g., [23]). Expectation refers to the degree of anticipating or being taken unaware and intensity refers to how far a person is away from a state of pure, cool rationality.

Although for many practical reasons arousal, valence and power dimensions have been assumed to be independent from each other, Dietz and Lang reported that these emotion dimensions are correlated [24]. Therefore, how to better exploit and model the correlations between emotion dimensions should be investigated further. Kaernbach in [25] argues that dimensional models may obfuscate the mechanisms underlying the genesis of emotions by drawing a parallel to dimensions of gustatory experience. Overall, he claims that too little is known about basic emotional processing in order to model it with a number of dimensions.

Scherer and colleagues introduced another set of psychological models, referred to as componential models of emotion, which are based on the appraisal theory [22,11,17]. In the appraisal-based approach emotions are generated through continuous, recursive subjective evaluation of both our own internal state and the state of the outside world (relevant concerns/needs) [22,26,11,17]. Despite pioneering efforts of Scherer and colleagues (e.g., [27]), how to use the appraisal-based approach for automatic measurement of affect is an open research question as this approach requires complex, multicomponential and sophisticated measurements of change. In a more recent article, Mortillaro et al. [28] suggest various ways to link automatic emotion recognition and appraisal models of emotion. This link aims to enable the addition of contextual information into automatic emotion recognizers, and enrich their interpretation capability in terms of multiple scales (more sensitive representation) and continuous dimensions (richer representation). Essentially, their approach suggests the use of continuous appraisal variables as an intermediate layer between the input expressive features and the emotion label output. This approach divides the emotion recognition process into two mappings: expressive features to appraisal variables, and appraisal variables to emotion label. Mortillaro et al. argue that the use of the componential theoretical perspective, with its set of continuous appraisal variables, can provide a number of benefits for automatic emotion recognition: 1) enhanced recognition of mixed (or multiple) emotions, mediated by the use of appraisal variables as outputs of the emotion recognition process; 2) ability to integrate the crucial, but currently missing, contextual information into automatic recognizers; and 3) facilitation of a finer level of interpretation (in terms of intensity and subtlety of the emotions predicted). Future research might be able to show whether these arguments are feasible or not.

2.2. The social sciences approach applied in engineering

Despite the existence of various other models, the categorical and dimensional approaches are the most commonly used models for automatic analysis and prediction of affect in continuous input.

In the case of the categorical approach there is some hierarchy from main categories such as positive, neutral, negative, or the big *n* emotions, e.g., anger, fear, sadness, joy, etc. to sub-categories modeling different shades of the main categories. These categories can be thought of as fixed or graded (e.g., weak, medium, strong), or as pure or mixed, or sometimes even antagonistic, if for instance a mixture of anger/joy/irony is being observed.

In the case of the dimensional approach, the foremost questions are how many and which dimensions we should base our analysis on. Traditionally, arousal and valence are modeled, with or without a third category power/dominance/control. Dimensional modeling can be more or less continuous, and if we assume more than one dimension, it automatically results in blended representation of emotions in the *n*-dimensional space. If emotion is conceptualized in a broader meaning, most likely, some other dimensions representing social, interactive behavior can be modeled as well [29].

Irrespective of strong beliefs in the one or the other type of modeling, in practice, categories can always be mapped onto dimensions and vice versa, albeit not necessarily lossless. It has been reported that discretization of the continuous dimensions for certain data sets results in an unbalanced range for some classes (e.g., [30]). This is attributed to the fact that the available data sets contain more neutral or non-emotional data than emotional data. Overall, other ways for discretization should be investigated. Is this suggestion at odds with the continuous affect representation? Probably not. Both representation models may be utilized to create a hierarchical structure that will involve granularities at multiple levels, e.g., from coarse categories towards a fine-grained continuous structure.

We will next look at how affect analysis is achieved in continuous input, which representation is used, and how processing is done.

3. Affect analysis in continuous input: a generalized view

3.1. Continuous vs. non-continuous input

One of the major research problems for creating automatic analyzers is to train them to predict affective behavior *correctly* (to match human

observers labeling), *robustly* (not too sensitive to changing conditions), and *quickly* (real-time processing and responsiveness). The nature of input data being continuous or non-continuous is of utmost importance when working towards achieving this goal.

Automatic analysis in continuous input differs from that of non-continuous in a number of ways. In non-continuous input, data is usually segmented such that it is constrained to contain one affective event (e.g., head nod/shake), expression (e.g., smile) or affective state (e.g., pain) with beginning and end. In continuous input instead the automatic analyzer ultimately needs to determine the starting and ending times of affective events, expressions and affective states (the segmentation). Segmenting multimodal data in an appropriate and meaningful way is directly related to determining the duration of an affective event. Determining analysis duration can also be seen as determining the window size [8] or the unit of analysis, as it is referred to by the automatic speech analysis community [31], to achieve optimal affect prediction. This is one of the issues that the existing literature does not provide a unique answer to. The current solution is to employ various window sizes depending on the modality. On the one hand achieving real-time affect prediction requires a small window size to be used for analysis (i.e., a few seconds, e.g., [32]), while on the other hand obtaining a reliable prediction accuracy requires longer-term monitoring [33,34]. Chanel et al. [32] reported large differences in accuracy between the EEG and peripheral features which may be due to the fact that the 8 s length of trials may be too short for a complete activation of peripheral signals while it may be sufficient for EEG signals.

Segmenting the affective data is also related to finding a condition against which changes in measured signals can be compared, i.e., determining the baseline. For bio signals, the baseline problem refers to finding a state of calmness [35]. For the visual modality the aim is to find a frame in which the subject is expressionless and against which changes in subject's motion, pose, and appearance can be compared. The current solution is to either segment the recordings (e.g., [36,37]), or to assume that at system startup the user is in a relatively expressionless (neutral) state (e.g., [5]).

The segmentation problem from speech is often addressed by using a unit of analysis, defined as a speech episode which has been segmented and often, but not necessarily, stored as a single speech file based on some criteria [31]. Often, this is a so-called 'turn' which starts when a speaker starts speaking, and ends when the speaker stops speaking in a conversation, or when the dialogue partner takes the turn. While defining a turn in interaction is actually very difficult, this is a well-defined and straightforward measure that is easily obtainable for acted data and for short dialogue-based conversations. However, in unconstrained settings such 'turns' can be very long — or even as short as "mm-hmm". In such cases, either some intuitive notion of emotional unit is used as a criterion, or some objective measure such as silences longer than, e.g., 0.5 s, or 1 s. However, such prosodic units may be too long and the marking of specific and shorter emotional episodes might be smeared, resulting in sub-optimal recognition performance. Two strategies, coping with these problems, can be observed. One defines 'technical units' such as frames, time slices, or proportions of longer units. For instance, the unit can be subdivided into three parts of equal length. The other one defines meaningful units with varying length such as syllables, words, phrases, i.e., chunks that are linguistically and by that, semantically, well-defined.

Overall, the challenge for future research is to find an appropriate unit of analysis. This is of utmost importance for three reasons [31]: firstly, emotionally consistent units will favor an optimal classification performance. Secondly, incremental processing, i.e., providing an estimate of the emotion before a longer utterance is finished, will often be necessary in real-life conversational systems, to enable reasonably fast system reactions [38]. And finally, for multimodal processing, an alignment of the different streams of information speech, facial gestures, bio signals, etc. has to be found anyway. In the long run, ad hoc segmentation strategies will have to be replaced by the alignment of meaningful units for each modality, taking into account their relevance for the specific setting.

3.2. A generalized view

In the context of this paper, automatic detection or prediction in continuous input refers to acquiring and processing long unsegmented recordings for detection of an affective state or event (e.g., pain) or continuous prediction of an affect dimension (e.g., valence, arousal, power). The emphasis here is on developments in the last couple of years (2009–2012) and on identified challenges and future prospects. Therefore, the focus of this paper is on affect analyzers created for somewhat realistic settings and trained/tested with naturalistic and spontaneous data. Details on 'earlier' databases, automatic analyzers and various related work can be found in [39,8,31], and [40].

The current trends in affect analysis in continuous input are discussed under the following headings: modalities and cues, data acquisition and annotation, automatic analysis and prediction, frameworks and tools, and context and applications.

4. Modalities and cues

An individual's inner emotional state may become apparent by subjective experiences (how the person feels), internal/inward expressions (bio signals), and external/outward expressions (vocal/visual signals). The research problem here is to identify which modalities and cues carry what kind of affective information to be used for continuous, fast and efficient affect analysis. As most of the affective science researchers assumed that the subject's subjective experience is directly related to the internal and external expressions, the current solution has been to focus on the analysis of the cues manifested both in the internal and the external communicative modalities. The internal communicative cues are grouped under bio signals and motion capture signals, and the external ones are grouped under visual, vocal, and textual cues.

4.1. Bio signals

Bio signals are multichannel recordings from both the central and the autonomic nervous systems. The bio signals used for automatic measurement of affect are galvanic skin response that is used as an indication of a person's level of physiological arousal [41], electromyography (the electrical potential mostly originated from muscular cell activities), that is correlated with negatively valenced emotions [42], heart rate that increases with negatively valenced emotions such as fear, heart rate variability that indicates a state of relaxation or mental stress, and respiration rate (how deep and fast the breath is) that becomes irregular with more aroused emotions like anger or fear [41,42].

Measurements recorded over various parts of the brain including the amygdala potentially enable observation of the emotions felt [43]. 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. It is also possible to observe the differences between positive and negative emotional stimuli from asymmetrical brain activity [44]. Despite the amygdala's role and importance in the brain activities due to emotions, the amygdala's deep location in the brain poses a challenge for obtaining valuable information via the EEG measurements. A number of studies also suggest that there exists a correlation between increased blood perfusion in the orbital muscles and stress levels for human beings. This periorbital perfusion can be quantified through the processing of thermal imagery (e.g., [45]).

Compared to sensors that sense vocal and visual signals, bio sensors are usually perceived as being invasive and cumbersome. More recently such aspects are mitigated by creating wearable sensors that are wireless and miniaturized (e.g., the BodyANT sensor [46] and Emotiv's

Epoc neuroheadset [47]). Despite such advances in related fields, obtaining accurate measurements from bio sensors is still affected by human physical activity (e.g., walking, bending, running).

4.2. Nonverbal and verbal vocal cues

Vocal signals convey affective information through explicit linguistic messages and implicit acoustic and prosodic messages that reflect the way the words are spoken. There exist a number of works focusing on how to map vocal expressions to dimensional models. Cowie et al. used valence-activation space, similar to valence-arousal space, to model and assess affect from speech [48,49]. Scherer and colleagues have also proposed how to judge emotional effects on vocal expression, using the appraisal-based theory [11]. In terms of affect recognition from vocal signals the most reliable finding is that pitch appears to be an index into arousal [50]. Another well-accepted finding is that mean of the fundamental frequency (F0), mean intensity, speech rate, as well as pitch range [51], 'blaring' timbre [52] and high-frequency energy [53] are positively correlated with the arousal dimension. Shorter pauses and inter-breath stretches are indicative of higher activation [54].

There is relatively less evidence on the relationship between certain acoustic parameters and other affect dimensions such as valence and power. The acoustic parameters that are correlated with arousal seem to also contain information for the perception of power dimension. Additionally, vowel duration and power dimension in general, and lower F0 and high power in particular, appear to have correlations. Positive valence seems to correspond to a faster speaking rate, less high-frequency energy, low pitch and large pitch range [53] and longer vowel durations. A detailed literature summary on these can be found in [55] and [56].

As for linguistic analysis of spoken or written words and phrases, it is generally reported that these are better suited for analysis of valence than for arousal [57]. In [57], the authors report that keywords such as *again*, *assertive*, and *very* are correlated with the arousal dimension, whereas *good*, *great*, and *lovely* are examples of such correlations with the valence dimension. In general, it seems logical that certain groups such as appraisal words or swear words are good indicators for the emotion at hand. Nonverbal vocalization is often handled in one string together with linguistic entities such as in 'It takes my mind off <laughs>' [58]. In vector space representation <laughs> would then be handled as a feature, just as any other linguistic term. This allows for mixed N-Grams of linguistic and nonverbal entities, which can be very meaningful, such as the bigrams 'no <laughs>' or 'no <sighs>'. There seems to be evidence indicating that laughter occurs more frequently in case of positive valence [59] whereas polarity seems to be encoded in the final position of the text [60]. However, to date there seems to be limited agreement on which part of speech class such as noun, verb, adjective, adverb, etc. [61] or class combinations such as adjective–adverb [62] or adjective–verb–adverb [63], are best suited to reflect affect in text.

4.3. Visual signals

Facial actions (e.g., pulling eyebrows up) and facial expressions (e.g., producing a smile) [64], and to a much lesser extent bodily postures (e.g., backwards head bend and arms raised forwards and upwards) and gestures (e.g., head nod) [65], form the widely known and used visual signals for automatic affect analysis and synthesis. Compared to the facial expression literature, attempts for recognizing affective body movements are few and efforts are mostly on the analysis of posed bodily expression data [66].

Detection of affect from bodily expressions is mainly based on categorical representation of affect. The categories happy, sad, and angry appear to be more distinctive in motion than categories such as pride and disgust. Darwin suggested that in anger, for instance, among other behaviors, the whole body trembles, the head is erect, the chest is well expanded, feet are firmly on the ground, elbows are squared [67,68]. Wallbot also analyzed emotional displays by actors and concluded that discrete emotional states can be recognized from body movements and postures. Analysis of the drinking and knocking arm movements showed that discrete affective states are aligned with the arousal–pleasure space [69]; and arousal was found to be highly correlated with velocity, acceleration, and jerk of the movement. To date, the bodily cues that have been more extensively considered for affect detection are static postural configurations of head, arms, and legs [70,71], static configurations and temporal segments [72], dynamic hand and arm movements [68], head movements (e.g., position and rotation) [73] and head gestures (e.g., head nods and shakes) [74,75]. For two recent surveys on affective body expression perception and recognition, the reader is referred to [66] and [76].

A number of researchers have investigated how to map various visual signals onto emotion dimensions. For instance, [15] mapped the facial expressions to various positions on the 2D plane of arousal–valence, while [74] investigated the emotional and communicative significance of head nods and shakes in terms of arousal and valence dimensions, together with dimensional representation of *solidarity*, *antagonism* and *agreement*. Their findings suggest that both head nods and shakes clearly carry information about arousal. However, their significance for evaluating the valence dimension is less clear [74]. In particular, the contribution of the head nods for valence evaluation appears to be more complicated than head shakes, e.g., "they understand what you say, and they care about it, but they do not like it".

Studies have shown that there is a relationship between the notion of approach/avoidance via the body movements and affective experiences [77,78], e.g., as a feedback of positively and negatively valenced emotions [79], postural leaning forwards and backwards in response to affective pictures [80], etc.

Gait, in the context of perception and detection, refers to a person's individual walking style. Therefore, gait is also a source of dynamic information by definition and has recently been exploited for emotion perception and recognition [81,82]. How people perceive the expression of emotional states based on the observation of different styles of locomotion has also been investigated in [83] by generating animation of a virtual character. The investigated characteristics were the head/torso inclination, the walking speed, and the viewing angle. Participants perceived distinct states of arousal and valence independent of the viewing angle.

The emerging trend in visual signal analysis is using cameras and sensors based on depth information (e.g., Microsoft Kinect). Such sensors provide quick solutions to problems pertaining to common vision-based analysis approaches (e.g., segmentation of the human body, etc.). However, there are range- and calibration-related issues that need to be solved prior to using them for a wider range of applications (e.g., analysis of face and facial features).

4.4. Motion capture signals

Although in a stricter sense not seen as part of the visual modality, motion capture systems have also been utilized for recording the relationship between affect dimensions and facial feature information [84], affect dimensions and body posture (e.g., [71,85]), and affect dimensions and body language (e.g., [86]. For instance, Kleinsmith et al. [85] identified that scaling, arousal, valence and action tendency were the affective dimensions used by human observers when discriminating between various body postures. Metallinou et al. found that whole body MoCaP features are not sufficient for predicting valence in the context of improvised dyadic acting [86]. They also found that increase in activation is often displayed by more walking and more approach behaviors towards the interlocutor.

4.5. Discussion

Exploring which modality is more closely correlated with which affective state or affect dimension is still being investigated. While visual cues and textual cues appear to be better for interpreting valence, vocal cues seem to be better for interpreting arousal [8]. Speech in general is reported to be less affected by the power dimension [15], than the arousal dimension. A thorough comparison between all modalities would indeed provide a better understanding of which emotion dimensions are better predicted from which modalities or cues.

Additionally, there are a number of challenges that still need to be investigated when input from multiple expressive channels is available. Firstly, the affective message conveyed by different modalities might be congruent (i.e., agreeing) or incongruent (i.e., disagreeing), depending on the context (e.g., feeling angry and not expressing it outwardly). Congruency has been investigated to some extent in multimodal expression and perception of categorical emotions [87,88]. A second challenge is cross-modal interactions between different cues and modalities. A number of studies investigated the combined perception of human facial and bodily expressions [87], and a number of findings illustrate the importance of emotional whole-body expressions in communication when viewed in combination with facial expressions and emotional voices [88]. Overall, further research is needed in multicue and multimodal affect expression and perception in order to explore how cross-modal interactions affect categorical and dimensional affect modeling and recognition.

5. Data

5.1. Data acquisition and annotation

In the affective computing research field, the data acquisition process needs to consider application domain, subjects (in terms of age, gender and cultural background), modalities, number and type of affective states, and type of data to be recorded. The recorded data type can fall into one of the following categories: acted (posed), re-acted (induced via clips) and inter-acted (occurring during an interaction).

Similarly to the field of sociolinguistics, when acquiring and annotating data, one of the problems faced by researchers in affective computing field is the Observer's Paradox, identified by Labov as a situation where data gathering is influenced by the presence of the experimenter [89]. Yet, acquiring affective data without subjects' knowledge is strongly discouraged and the current trend is to record naturalistic (spontaneous) data in more constrained conditions such as an interview or a recall paradigm (e.g., recall of past emotional life episodes [32]) or interaction (e.g., [23]) setting, where subjects are still aware of placement of the sensors and their locations.

Currently, there exist a number of annotation tools with different capabilities, used for different purposes in different context. Of these, ELAN is widely used for annotating affective behavior using categorical descriptions. It allows continuous multimedia annotation with separate annotation layers that contain linguistically relevant descriptions of events that occur within the track [90]. The audio and/or video is displayed together with these annotations. ANVIL is another widely used video annotation tool introduced in 2001 [91]. The tool has been developed to enable the users to use their own coding scheme, to save data in XML format, and to see color-coded elements on multiple tracks in time-alignment. The latest version, *ANVIL 5*, contains a number of features such as cross-level links, non-temporal objects, time-point tracks, coding agreement analysis, and a project tool for managing whole corpora of annotation files. An important ANVIL feature to come is the ability to import ELAN files.

The FEELtrace annotation tool is used for annotating expressions displayed via vocal and visual signals with continuous traces (impressions) in the dimensional space. FEELtrace allows coders watch the audiovisual recordings and move their cursor, within the 2-dimensional emotion space of valence and arousal confined to $[-1, +1]$, to rate their impression about the emotional state of the subject [48]. The motivation behind this interface is to provide some visual feedback to the annotator in terms of where he is on the given scale without interfering with his attention to the material being rated. More recently, 'General trace' (Gtrace) has been introduced as a descendant of FEELtrace, to let people create their own scales with minimum effort along a chosen communicative or affective dimension [92]. Gtrace allows the annotator watch the video to be rated and simultaneously see a cursor that they can manipulate. The cursor is in the form of a colored disk that can be moved along to the left and to the right.

For annotating the internal expressions (bio signals), the level of valence and arousal is usually extracted from subjective experiences or subjects' own responses (e.g., [43,93]) due to the fact that feelings induced by an image or sound can be very different from subject to subject [94]. The Self Assessment Mannequin (SAM) [95] is the most widely used means for self assessment. Another tool called *the motion slider* allows the collection of self-reported valence information from subjects while they interact with a system [96].

Another major challenge in affective data annotation is the fact that there is no coding scheme that is agreed upon and used by all researchers in the field that can accommodate all possible communicative cues and modalities. Development of an easy to use, unambiguous and intuitive annotation scheme that is able to incorporate inter-observer agreement levels will indeed ease the heavy burden of the annotation task.

When discretized dimensional annotation is adopted (as opposed to the continuous one), researchers seem to use different intensity levels: either a ten-point Likert scale (e.g., 0 — low arousal, 9 — high arousal), or an arbitrary range, e.g., between -50 and $+50$ [97], or between -1 and $+1$, divided into a number of levels [8]. The final annotation is usually calculated as the mean of the observers' ratings. Other variants such as the 'median' can constitute an alternative [94]. In addition, individual weighting of evaluators can be obtained and used by the so-called evaluator weighted estimator (EWE), as described in [98].

Obtaining high inter-observer agreement is a challenge in affect data annotation, especially when the continuous dimensional approach is adopted. To date, researchers have mostly chosen to use self-assessments (e.g., [42]) or the mean within a predefined range of values of the observers' ratings (e.g., [71]). Other methods, that take into account agreement and correlation measures, have also been proposed (e.g., [99,36]). Overall, deriving appropriate ground truth from both discretized and continuous dimensional annotations, modeling inter-observer agreement levels within automatic affect analyzers, and finding which signals better correlate with self assessments and which ones better correlate with independent observer assessments remain as challenging issues in the field.

Another challenging aspect of data annotation is the fact that some affective states (e.g., pain) are difficult to assess and manage. For instance, the affective state pain is a subjective phenomenon and is typically measured by self-report [100]. Needless to say self-report measures have their limitations: they may be inconsistent, variable, and depend on past experience. The quality of self-report labels can be improved by rigorously training the participants in the labeling scheme adopted (e.g., [101]).

Recent studies provide a number of suggestions that need to be explored further. One suggestion is that employing multiple sources of affective ground truth information (triangulation) can aid creating better automatic affect recognizers [101]. Another suggestion is that for automatic detection of long behavioral states (e.g., pain or depression) instead of frame-by-frame behavioral coding somewhat coarse ground truth could be sufficient [100].

5.2. Databases

The research problem in affective database creation for continuous analysis is creating naturalistic settings that will encourage the

participation and responsiveness of the subjects for relatively longer duration. The current solution to this problem has been to define a realistic context that will induce naturalistic affective behavior over the course of time.

Another good source for such data has been collaboration with clinical researchers and obtaining access to clinical data. However, due to high level of sensitivity and privacy involved in these recordings, such data is usually not publicly available for other researchers. Despite such challenges, there is still a growing body of databases that contain naturalistic multimodal and continuous (unsegmented) data, labeled continuously either in terms of discrete categories or along the emotion dimensions, and made publicly available for research purposes.

The new trend in affect database creation is to provide baseline detection and prediction results together with the created databases. In this way other researchers using the available data are able to compare their methods to the baseline methods and results.

We will first look at the databases created for continuous affect detection and annotated in terms of affect categories. The UNBC-McMaster database [102] contains patients with shoulder injuries portraying real or spontaneous pain. The Multi-Modal Affective Database for Affect Recognition and Implicit Tagging (MAHNOB-HCI) [103] is a collection of various modalities recorded in a synchronized manner. The recorded cues include multicamera video of face, head, speech, eye gaze, pupil size, EEG, ECG, GSR, respiration amplitude and skin temperature. The authors also provide emotion recognition and implicit tagging results from different modalities in order to set a baseline result for researchers who are going to use the database in the future. The QMUL-UT EEG Dataset contains multimodal affective data, including EEG and physiological signals such as EOG, GSR, heart rate, temperature and respiration, of subjects watching video sequences depicting events, followed by either a matching or a non-matching tag [104].

The Database for Emotion Analysis using Physiological Signals (DEAP) [105] contains spontaneous physiological signal recordings and face videos of participants watching and rating their emotional response to 40 music videos along the scales of arousal, valence, and dominance, as well as their liking of and familiarity with the videos. The authors also provide classification results using various features (from the EEG, peripheral and MCA modalities) and combination of features, and performing single-trial (single-participant) classification (for the scales of arousal, valence and liking). They report that modalities appear to perform moderately complementary, where EEG performs best for arousal, peripheral for valence and MCA for liking.

Next we will look at the databases created for continuous affect analysis that contain unsegmented affect data and as well as data annotated in terms of affect dimensions. The SEMAINE database [106] contains annotated multimodal recordings of emotionally colored conversations between a person and a limited agent. It has high quality audiovisual recordings of 150 participants interacting with different configurations of a Sensitive Artificial Listener (SAL) agent. The database contains 959 conversations (approximately 5 min, each) rated by 6–8 raters along 27 associated categories or dimensions.

The Belfast Induced Natural Emotion Database [107] contains recordings of mild to moderate emotionally colored responses to a series of laboratory based emotion induction tasks and annotations in terms of self-reports, continuous trace-style ratings of dimensions, and several other characterizing parameters.

The new emerging trend in continuous affective data acquisition is to focus on multimodal and multispeaker dyadic interactions rather than human–computer or human–technology interaction (e.g., [86]). The UCS CreativeIT database was created with the aim of studying affective communication and interaction between humans [108]. It contains improvisation data from pairs of theater actors that were recorded via cameras, Motion Capture (MoCap) markers placed over their full body and close talking microphones. The long unsegmented recordings lasted 2–8 min, and were annotated using FeelTrace along the dimensions of valence (positive–negative), activation (excited–calm) and dominance (dominant–submissive). A multimodal database for mimicry analysis has been introduced in [109]. The database was recorded using 18 synchronized audio and video sensors and two dyadic interaction settings where participants had a discussion on a political topic, and a role-playing game. The database contains 54 recordings from 40 participants and 3 confederates (26% female and 95% southern European). Metadata is also made available together with the recordings (e.g., dialogue acts, turn-taking, etc.).

More recent naturalistic affect databases can be found in the Special Issue of IEEE Tran. on Affective Computing [110].

6. Automatic analysis and prediction

The challenges in working with continuous input bring along a number of relevant sub-problems, namely, feature extraction, finding an optimal set of features, creating prediction methods that can handle continuous input, and training automatic predictors that can generalize well. The feature extraction techniques used for each communicative source are similar to the previous works (reviewed in [39]) in the field. For further details on how features are extracted for each communicative modality, and how multicue and multimodal fusion is achieved for affect analysis purposes see [8,39,40].

Finding the optimal set of features is a challenge for automatic analyzers as it directly affects their detection and prediction accuracy. There exist a number of studies focusing on finding the most appropriate features for predicting arousal, valence and/or power dimensions from speech (e.g., [111,112]). Similarly, some studies investigated how cue masking or filtering affects the perception of emotion. This was done for the vocal modality by separating the semantic content from the acoustic channel [113]. For the visual modality Cohn and colleagues investigated how cue masking affects behavior perception and generation by separating appearance and motion, and how manipulating temporal dynamics affects behavioral event production of the conversing partners. They found that head nodding is regulated by dynamics rather than by the partner's evident gender [100]. Such findings have implications for meaningful interpretation of sensor data and should be carefully considered when designing automatic affect analyzers and interpreters.

Generalization capability of automatic affect analyzers across subjects remains as a research problem in the field. Kulic and Croft [93] reported that for bio-signal-based affect measurement subjects seem to vary not only in terms of response amplitude and duration, but for some modalities, a number of subjects show no response at all. This makes generalization over unseen subjects a very difficult problem. A common way of measuring affect from bio signals is doing it for each participant separately without computing a baseline, e.g., [32]. The current solution provided by recent works on automatic affect prediction from vocal or visual cues is to compare subject-dependent vs. subject-independent prediction results (e.g., [114]) in order to obtain better insight into this issue.

In the following sections we review the automatic analysis and prediction techniques employed for each modality by mostly focusing on a number of representative works introduced after 2009. Details on single-modal and multimodal systems focusing on dimensional affect analysis can be found in [8]. For details on earlier works (e.g., [115–120]) the readers are referred to [8] and [36].

6.1. Bio signals

Prior to extracting features, affect recognition systems that use bio-signals as input usually pre-process signals to remove noise. 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 [121]. Following the preprocessing stage, there are various alternatives for

feature extraction, e.g., mean, standard deviation, mean of the absolute values of the first differences, etc. (e.g., [32,122]).

Automatic affect prediction from bio signals is usually done by reducing the prediction problem to a two-class classification problem, e.g., arousal vs. non-arousal and valence vs. non-valence [123]. New trends such as multi-stage classification and identification of boundaries within the continuous emotion space have also started to emerge. For instance, Khosrowabadi et al. present in [124] an EEG-based emotion recognition system using a self-organizing map to identify the boundaries (threshold levels) between separable regions of the arousal and valence dimensions, approached as four emotion categories. Frantzidis et al. [125] recorded bio signals while subjects viewed affective pictures. The recorded bio signals were first classified along the valence dimension, and together with gender information were input to a second layer distance classifier that classifies the data into high and low arousal [125].

With the availability of low-cost EEG head-sets, such as the Emotiv EPOC [47], the use of neurophysiological signals for online affect and mental state detection has been gaining momentum. Neurophysiological signals are also being used to investigate brain activity associated with the expression of facial actions, and recognize the facial action displayed (e.g., blink) [126]. For various studies using neurophysiological signals for affective data collection and analysis, and relevant applications on affective Brain–Computer Interaction, see [127] and [128].

6.2. Vocal signals and textual cues

Similarly to the affect recognition from bio signals, the most commonly employed strategy in automatic dimensional affect prediction from vocal signals is to reduce the prediction problem to a two-class classification problem (positive vs. negative or active vs. passive classification; e.g., [129]), a three-class classification problem (lower, middle and higher) (e.g., [30]) or a four-class classification problem (classification into the quadrants of 2D arousal–valence (A–V) space; e.g., [57]).

As far as continuous affect prediction without quantization is concerned, there exist a number of methods that deal exclusively with speech (i.e., [57,130,131]). The work by Wöllmer et al. uses the SAL Database [132] and Long Short-Term Memory (LSTM) neural networks and Support Vector Machines for Regression (SVR) [130]. The work presented in [57] utilizes a hierarchical dynamic Bayesian network combined with Bidirectional Long-Short Term Recurrent Neural Networks (BLSTM-RNN) performing regression and quantizing the results into four quadrants after training. Grimm and Kroschel use the "Vera am Mittag" (VAM) German audio-visual emotional speech database [99] and SVRs, and compare their performance to that of the distance-based fuzzy k-Nearest Neighbor and rule-based fuzzy-logic estimators [131]. The work by Espinosa et al. uses the VAM database [99] and examines the importance of different groups of speech acoustic features in the estimation of the three-dimensional continuous model of emotions (PAD) [21].

Looking at textual cues, be they from the speech signal subsequent to speech recognition, or directly from text such as in text-based chat or posts [133], a variety of approaches are employed. These approaches include calculating affective salience from single words' or N-Grams', obtaining vector space representations based on bag-of-words, or using combination of these approaches by using bags-of-(character)-N-Grams. Other approaches include tailored machine learning techniques such as string kernels for Support Vector Machines, or knowledge source exploitation [134] such as diverse affective word lists, EmotiNet [135], Concept Net, General Inquirer or WordNet(-Affect) and alike [136]. Apart from direct exploitation of the word string, and in addition to using the original word string, re-tagging by using parts of speech classes (e.g., verb, noun, adjective, etc.) were proven to be useful [137]. Additionally, nonverbal events such as laughter or sigh can be incorporated into the word string [58,138] as was shown above. Little influence was found by the fact that speech needed to be recognized automatically as long as the affect recognition is trained in matched condition on the speech recognizer output as compared to using manually transcribed text [139]. In line with this, little influence of stemming or stopping (i.e., clustering of morphological variants of a spoken or written term and exclusion of terms) is often reported, and word order dependencies are rather marginal [140]. This stems from the fact that affective key-phrases are typically rather short.

Discrete affective classes in text analysis are often very distinct from those in related modalities: Typical recent targets of interest comprise irony, sarcasm, satire, metaphor, or parody. Few works dealt with affect prediction from linguistic analysis of spoken language [58,136] or written language [141,140] (here also often referred to as sentiment analysis) in dimensional representation. These works show that vector space modeling seems among the most promising approaches: apart from giving good results, it allows for easy integration into an acoustic or multimodal feature vector. To date, continuous input handling for textual cues is virtually unexplored in the literature. The SEMAINE example [138] shows that in an online setting, where speech needs to be analyzed directly as it 'comes in', single-word-based analysis can be the most practical solution.

6.3. Visual signals

There exists an extensive literature for face and body feature extraction, tracking and gesture recognition from video sequences. In the context of affect sensing, we only briefly summarize the existing trends. The facial feature extraction techniques, used for categorical and dimensional affect analysis from the visual modality, can be categorized under two categories [64]: 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, and change in texture are measured. Hybrid approaches explore the combination of these two.

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 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). For details see the relevant survey papers [142,143].

The most commonly employed strategy in automatic dimensional affect prediction from visual signals is to reduce the prediction problem to a two-class classification problem (positive vs. negative or active vs. passive classification; e.g., [120,144]) or a four-class classification problem (classification into the quadrants of 2D A–V space; e. g., [145,146]).

Touching upon the segmentation problem from video, a technique to extract emotional segments from video based on the pleasure–arousal–dominance (P–A–D) model, assuming independency between the dimensions, is introduced in [147]. Nicolau et al. also focus on automatically segmenting emotional clips from long audiovisual interactions in [148]. Overall, however, there is no agreement on (i) whether continuous prediction should be done without segmentation, and (ii) whether segmenting videos into shorter clips is useful for continuous emotion prediction.

Currently, there are also a number of works focusing on dimensional and continuous prediction of emotions from the visual modality [75,149,114]. Kipp and Martin in [149] investigate how basic gestural form features (e.g., preference for using left/right hand, hand shape, palm orientation, etc.) are related to the single PAD dimensions of emotion, without performing automatic prediction. The work by [75] focuses on dimensional prediction of emotions from spontaneous

conversational head gestures by mapping the amount and direction of head motion, and occurrences of head nods and shakes into arousal, expectation, intensity, power and valence level of the observed subject using SVRs. The work by Nicolaou et al. focuses on dimensional and continuous prediction of emotions from naturalistic facial expressions within an Output-Associative Relevance Vector Machine (OA-RVM) regression framework by learning non-linear input and output dependencies inherent in the affective data [114]. Emotion recognition from gait has also been attempted. Janssen et al. in [81] focused on emotion recognition from human gait by means of kinetic and kinematic data using artificial neural nets. Their results showed that subject-independent emotion recognition from gait patterns is indeed possible.

6.4. Motion capture signals

Motion capture systems have mostly been utilized for recording the relationship between affect dimensions and facial feature information (e.g., [84]), and affect dimensions and body posture (e.g., [71,85]). In order to achieve this, geometrical features are extracted from the captured recordings with the aid of registration, a coordinate system definition, and subsequent computation of euclidean distances and relative positions and velocities. Higher-level abstraction is obtained by computing features such as orientation (e.g., orientation of shoulder axis) and distance (e.g., distance between left elbow and left shoulder) that appear to help in effectively discriminating between the (quantized) affect dimensions [71,85].

Karg et al. [82] focused on using both discrete affective states and affective dimensions for emotion modeling from motion capture data and showed that gait is a useful cue for the recognition of arousal and dominance dimensions.

Metallinou et al. in [86] focused on analyzing the vocal and body language behavior (via MoCap features) of pairs of actors improvising dyadic interactions. For each actor's recording, they computed the Spearman correlation coefficient between the mean annotation and the MLE curve. When predicting activation from visual and audiovisual cues, they obtained correlation histograms with relatively high values, for the dominance dimension values obtained were relatively lower, while for the valence dimension, the correlations were close to zero.

6.5. Thermal imaging signals

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 [45]. Some studies estimate differential images between the averaged neutral face/body and the expressive face/body [150] 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 [151]. Patterns of thermal variations for each individual affective state are also used. 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 [45].

Relatively few efforts have been reported on dimensional affect recognition from thermal imagery. Nhan and Chau [152] focus on recording the thermal infrared signals of the subjects stimulated with images from the International Affective Pictures System. They use the self-reported affect as ground truth and achieve high vs. low classification of arousal and valence using the time-frequency features derived from thermal infrared data. Merla and Romani utilized functional infrared imaging (fIR) to study the facial thermal signatures of three emotional conditions: stress, fear and pleasure arousal [153]. They reported that fIR can be reliably used to assess emotional arousal.

6.6. Modality fusion

Recent works focus on dimensional and continuous prediction of emotions from multiple modalities. In automatic affect prediction, feature-level fusion is obtained by concatenating all the features from multiple cues into one feature vector which is then fed into a machine learning technique (e.g., [86,36]. If the audio stream has a different frame rate with respect to the video stream (e.g., 50 Hz vs. 25 fps), during feature-level fusion each video feature vector is adapted to the audio stream (e.g., [36,37]).

In the decision-level fusion, the input coming from each modality and cue is modeled independently, and these single-cue and single-modal recognition results are combined in the end (see Fig. 3(a)). A representative approach is that of Gilroy et al. [154] where the affective behaviors of spectators are represented in terms of audio, video and attention events, and each input modality is first represented as a PAD vector, and then fused at the decision level using equal weights for all modalities. Unlike many other multimodal approaches (e.g., [120,155,116]), the ground truth in this work is obtained by measuring Galvanic Skin Response (GSR) as an independent measure of arousal.

Since humans display multi-modal expressions in a complementary and redundant manner, the assumption of conditional independence between modalities and cues in decision-level fusion can result in loss of information (i.e., loss of mutual correlation between the modalities). Model-level fusion has been adopted to mitigate the issues pertinent to decision-level fusion as it has the potential of capturing correlations and structures embedded in the continuous output of the classifiers or regressors from different sets of cues (see Fig. 3(a)).

Although automatic affect analyzers based on physiology end up using multiple signal sources, explicit fusion of multimodal data for continuous modeling of affect utilizing dimensional models of emotion is still relatively unexplored. A representative example to explicit fusion is the work of Eyben et al. [138] that proposed a string-based approach for fusing the behavioral events from visual and auditive modalities (i.e., facial action units, head nods and shakes, and verbal and nonverbal vocal cues) to predict human affect in a continuous dimensional space in terms of arousal, expectation, intensity, power and valence dimensions.

Overall, finding the type of annotation to be used as ground truth (i.e., self assessment, rater assessment or measurement-based assessment such as GSR) and the best way to fuse the modalities for continuous emotion prediction remain as open issues in the field.

An emerging trend in affective data fusion is called output-associative fusion (e.g., [36]). This fusion method capitalizes on the fact that the emotion dimensions (valence and arousal) are correlated [156,157,19], which has also been reported in [158]. In order to exploit these correlations and patterns, the output-associative fusion framework aims to learn the dependencies that exist among the predicted dimensional values (see Fig. 3(b)).

6.7. Classification schemes

The choice of classifier usually depends on the context and the application. Classification methods used for discrete affect detection and recognition include, among others, Support Vector Machines (SVM), Multi-Layer Perceptron networks, k-nearest neighbor classifiers, decision trees and their variations, Näive Bayes classifiers, Radial Basis Function networks, Linear Logistic Regression (LLR) [2], and Hidden Markov Models (HMM) and their variations (e.g., coupled HMM or asynchronous HMM) [30,120]. For instance, McDuff et al. in [144] focus on classification of self-report valences (positive, negative or

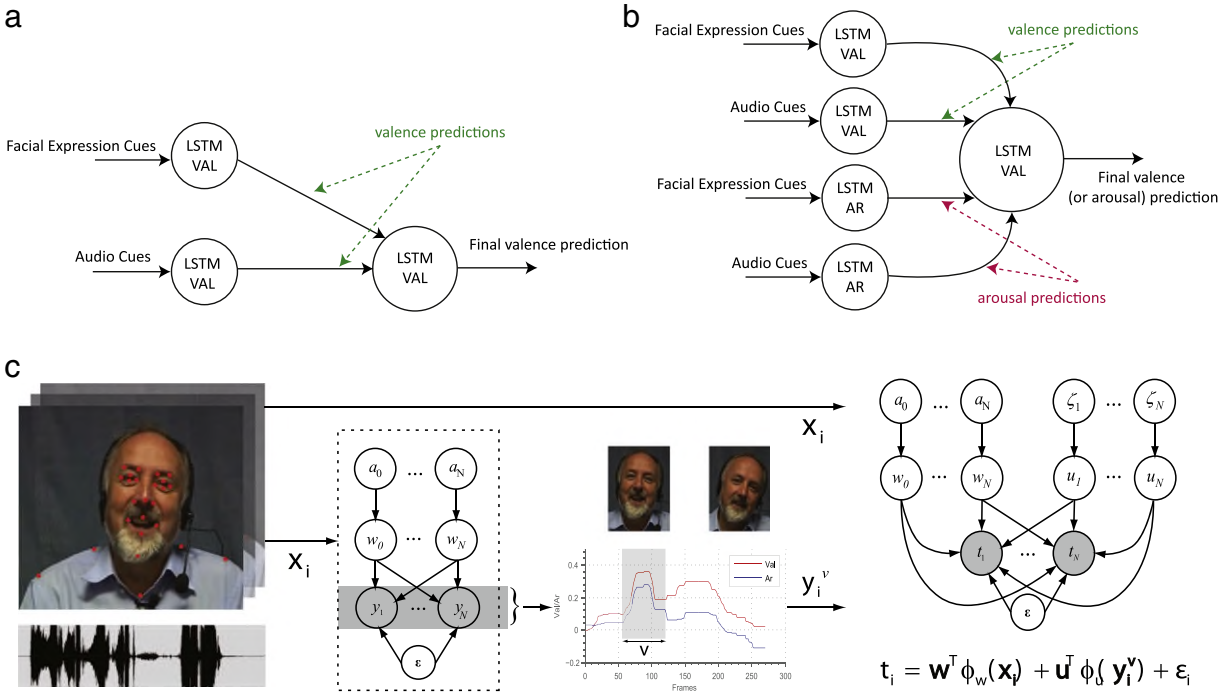

Fig. 3. Illustration of (a) model-level fusion (courtesy of [36]), (b) output-associative fusion (courtesy of [36] and (c) output-associative prediction (courtesy of [114])): (a) model-level fusion using facial expression and vocal cues for predicting valence. (b) Output-associative fusion combining valence and arousal values predicted from facial expression and vocal cues. (c) Output-associative prediction combines the valence and arousal values predicted from facial expression cues together with original input features for refining the prediction.

neutral) from spontaneous facial action units (in unsegmented videos) comparing various classifiers, namely, Support Vector Machines, Hidden Markov Models, Conditional Random Fields and Latent-Dynamic Conditional Random Fields. Various frameworks combining the benefits of multiple classifiers have also been proposed (e.g., combining Coupled HMM and SVM for classification [120], a multi-layer hybrid framework for classification [159]). Continuous affect measurements should be able to produce continuous values for the target dimensions. Some of the classification schemes that have been explored for this task are, SVR, Relevance Vector Machines (RVM), and LSTM Recurrent Neuronal Networks (RNN). Linear Discriminant Analysis (LDA), Conditional Random Fields (CRF) and SVM have been used for quantized dimensional affect recognition tasks (e.g., [130]). Overall for automatic affect analysis in continuous input there is no agreement on how to model dimensional affect space (continuous vs. quantized) and which classifier is better suited for automatic multimodal analysis of continuous affective input.

The design of emotion-specific prediction and classification schemes that can handle multimodal, continuous and spontaneous data is one of the most important issues in the field. In accordance with this, Kim and André proposed a novel scheme of emotion-specific multilevel dichotomous classification (EMDC) using the property of the dichotomous categorization in the 2D emotion model and the fact that arousal classification yields a higher correct classification ratio than valence classification or direct multiclass classification [160]. They apply this scheme on classification of four emotions (positive/high arousal, negative/high arousal, negative/low arousal and positive/low arousal) from physiological signals recorded while subjects were listening to music. An emotion-specific classification scheme is also introduced in [161]. The multi-layer hybrid framework (MF-Hybrid) comprises a temporal regression layer for predicting affect dimensions, a graphical model layer for modeling valence–arousal correlations, and a final classification and fusion layer exploiting informative statistics extracted from the lower layers. Creating such emotion-specific schemes for continuous prediction of emotions from other modalities should be investigated further.

An emerging trend in continuous affect prediction is the so-called *output-associative prediction* (e.g., [114]). This prediction method capitalizes on the fact that the emotion dimensions of valence and arousal are correlated [156,157,19]. In order to exploit these correlations and patterns, the output-associative prediction framework aims to learn the dependencies that exist among the predicted dimensional values. The framework consists of an initial layer of emotion dimension predictors, trained independently, that use the initial set of vocal or visual features as input. The framework then learns to map the outputs of these intermediate predictors onto a higher and final level of prediction by incorporating cross-dimensional output dependencies. Various predictors (BLSTM-NN, SVR, RVM, etc.) can be used as in the initial layer (a basis regressor) for this framework (see Fig. 3(c)).

6.8. Performance evaluation

Discrete affect detection methods that use continuous input obtain performance evaluation using the measures of detection rate, f1 measure and area under the ROC curve. Detection rate is either computed at the instance-level or segment-level. The instance-level refers to frame-level for video-based detection, and unit-level for audio-visual-based detection. It can be calculated per emotion category as the fraction of the number of segments correctly detected as that emotion category divided by the total number of segments available for that emotion category. F1 measure is the harmonic mean of the precision and recall (highest value being 1 and lowest value being 0). Area under the ROC curve accuracy is typically quantified as A^1 — the area under the receiver operating characteristics (ROC) curve. A^1 values can range between .5 (chance) and 1 (perfect agreement).

Finding optimal evaluation metrics for dimensional and continuous affect prediction remains an open research issue [8]. Pearson's correlation coefficient and the mean squared error (MSE) are the most

commonly used evaluation measure by related work in the literature (e.g., [130,131,116]). MSE evaluates the prediction by taking into account the squared error of the prediction from the ground truth. Let $\hat{\theta}$ be the prediction and θ be the ground truth. MSE is then defined as:

$$\mathrm{MSE} = \frac{1}{n}\sum_{f=1}^{n}\left(\hat{\theta}(f)-\theta(f)\right)^2 = \sigma_{\hat{\theta}}^2 + E\left(\left[\hat{\theta}-\theta\right]\right)^2. \quad (1)$$

As can be seen from the equation, MSE is the sum of the variance and the squared bias of the predictor, where E is the expected value operator and μ is the expected value. Therefore, the MSE provides an evaluation of the predictor based on its variance and bias. This also applies for other MSE-based metrics, such as the root mean squared error (RMSE), defined as:

$$RMSE = \sqrt{MSE}.$$

RMSE is also referred to as Mean Absolute Error (MAE) or Mean Linear Error (MLE). Some researchers opt for using the RMSE (e.g., [36]) as it is measured in the same units as the actual data at hand, as opposed to the squared units measuring MSE. MSE-based evaluation has been criticized for heavily weighting outliers [162]. Most importantly, it is unable to provide any structural information regarding how θ and $\hat{\theta}$ change together, i.e., the covariance of these values. Pearson's correlation coefficient (COR), employed for evaluating the prediction and ground truth, compensates for the latter, and is defined as follows:

$$\mathrm{COR}\left(\hat{\theta},\theta\right) = \frac{COV\left\{\hat{\theta},\theta\right\}}{\sigma_{\hat{\theta}}\sigma_{\theta}} = \frac{E\left[\left(\hat{\theta}-\mu_{\hat{\theta}}\right)(\theta-\mu_{\theta})\right]}{\sigma_{\hat{\theta}}\sigma_{\theta}} \quad (2)$$

where σ stands for the standard deviation, COV stands for the covariance while μ symbolizes the mean (expected value). COR provides an evaluation of the linear relationship between the prediction and the ground truth, and subsequently, an evaluation of whether the model has managed to capture linear structural patterns inhibited in the data at hand. As MSE cannot fully represent the performance of an automated system, correlation coefficient has been employed by several studies (e.g., [131,116]) together with MSE. In fact, several works entirely rely on this measure and do not report MLE or derivatives as this measure is not affected by delays that may exist between different labelers (i.e., temporal offset in labeling).

In addition to the two aforementioned metrics, another emotion-prediction-specific metric is proposed to obtain an agreement level of the prediction with the ground truth by assessing, e.g., the valence dimension, as being positive (+) or negative (−), and the arousal dimension, as being active (+) or passive (−) [36]. Based on this heuristic, a sign agreement metric (SAGR) is defined as follows:

$$\mathrm{SAGR} = \frac{1}{n}\sum_{f=1}^{n}\delta_{\left(sign\left(\hat{\theta}(f)\right),sign(\theta(f))\right)} \quad (3)$$

where δ is the Kronecker delta function, defined as:

$$\delta_{(a,b)} = \begin{cases} 1, & a = b \\ 0, & a \neq b. \end{cases} \quad (4)$$

The empirical evaluations of [36] showed that there is an inherent trade off involved in the optimization of the MSE (RMSE), COR, and SAGR metrics. By using all three metrics simultaneously it may be possible to attain a more detailed and complete evaluation of predictor vs. ground truth, i.e., (i) a variance-and-bias-based evaluation with RMSE — how much prediction and ground truth values vary, (ii) a structure-based evaluation with COR — how closely the prediction follows the structure of the ground truth, and (iii) emotion-prediction-specific evaluation with SAGR — how much prediction and ground truth agree on the positive vs. negative, and active vs. passive aspect of the exhibited expression.

An additional challenge is faced when attempting to measure temporal correctness. Temporal deviations come into play if the unit of analysis is not a single frame, and the task becomes that of 'emotion spotting' for certain emotions. This issue has not yet been sufficiently explored in the field.

7. Frameworks and tools

In the last five years, various research groups have created publicly available frameworks and tools to be used for researching dimensional and continuous analysis of emotions (e.g., [163–165]). Of these, the SEMAINE API introduced in [163] is an open source framework for building emotion-oriented systems, using standard representation formats and providing a Java and C++ wrapper around a message-oriented middleware. The openSMILE library is written in C++ and enables extraction of large audio feature spaces in real-time [164]. A middleware solution to aid the design and development of healthcare applications with affective information is introduced in [165].

The Computer Expression Recognition Toolbox (CERT) [166] is a software tool for fully automatic real-time facial expression recognition and is free for academic use. The tool outputs the locations of 10 facial features, estimates the 3-D head orientation (yaw, pitch, roll), detects 19 facial actions (AUs 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 18, 20, 23, 24, 25, and 26) and their intensity, and recognizes the 6 prototypical facial expressions (happiness, sadness, surprise, anger, disgust, and fear). CERT is able to analyze 320∗240 video sequences at 10 frames per second. The EyesWeb XMI Expressive Gesture Processing Library [167] is another useful tool for extracting features from affective body movement and gestures such as activity levels, spatial extent, symmetry and jerkiness of gestures, etc.

The interested readers are referred to the HUMAINE Portal [168] that is an excellent source of information about the latest frameworks and tools made available by researchers for various purposes.

8. Context and applications

Taking the automatic affect detection and prediction recognition systems into the wild and impacting the daily lives of ordinary people and users is the ultimate goal of affective computing. Context in that sense refers to the knowledge of who the subject is, where she is, what her current task is, and when the observed behavior has been shown. More recently, a number of works started exploring automatic analysis of affect in an application-dependent and context-specific manner in non-acted scenarios:

- human–computer interaction (e.g., Sensitive Talking Heads [169], Sensitive Artificial Listeners [5,7], spatial attention analysis [170], arts installations [171])
- human–robot interaction (e.g., humanoid robotics [172–174])
- clinical and biomedical studies (e.g., stress/pain monitoring [175–177], autism-related assistive technology [178])
- human behavior-analysis related applications (e.g., improving public speaking skills [14])
- learning and driving environments (e.g., measuring spontaneous facial expressions of children during problem solving tasks [179,180])
- developing affect-sensitive tutors (e.g., [180]) and episodic learning (e.g., [181])
- affect analysis in the car (e.g., [182])
- multimedia (e.g., video content representation and retrieval [183,184], personalized affective video retrieval [185,186])
- entertainment technology (e.g., gaming [187,188,173])

Defining and setting up a specific context enables designing automatic systems that are realistic, and are sensitive to a specific target

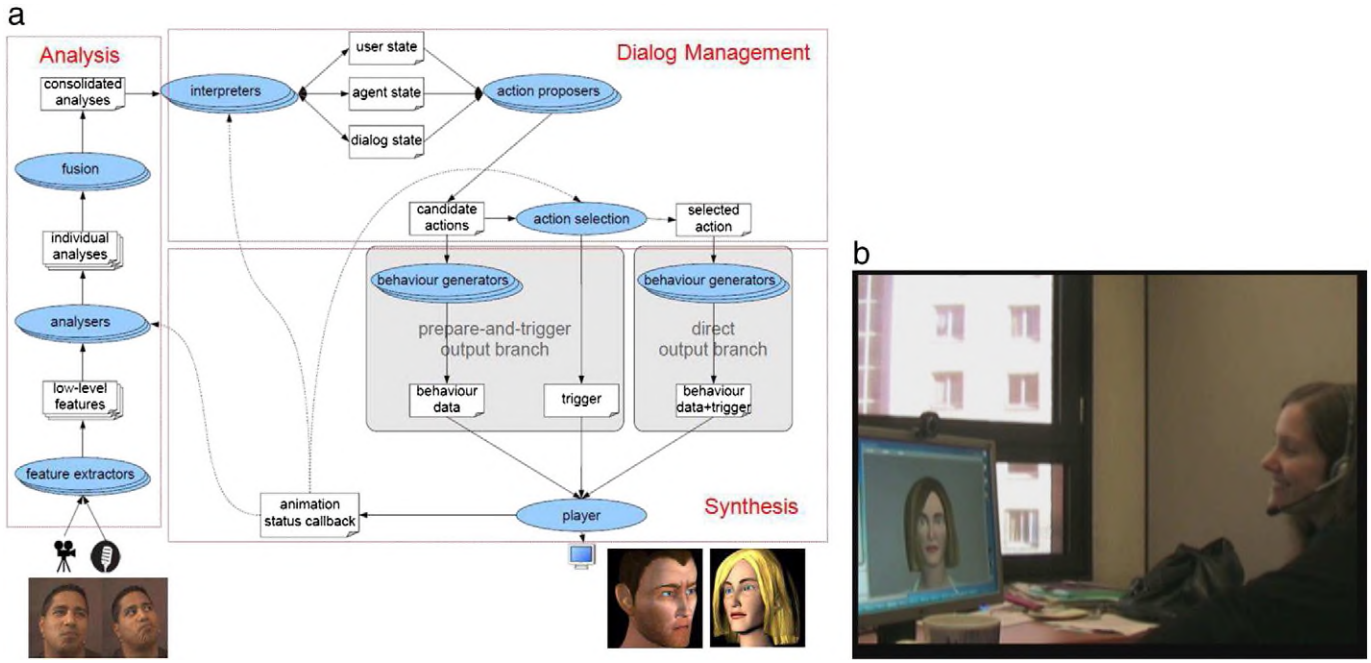

Fig. 4. Illustration of (a) the analysis, the dialogue management and the synthesis components of the fully Autonomous SAL system (courtesy of [5]), and (b) a human subject conversing with one of the SAL system characters (i.e., Poppy).

user group and target application. Defining a context potentially simplifies the problem of automatic analysis and recognition as the setup chosen may encourage the user to be in a controlled position (e.g., sitting in front of a monitor or standing in a predefined area), wearing specific clothes (e.g., wearing bright-colored t-shirts [173] or a motion capture suit [188]), etc. Overall, however, how to best incorporate and model context for affect analysis in continuous input from multiple cues and modalities needs to be explored further.

There are also spin off companies emerging out of collaborative research at well-known universities (e.g., Affectiva [189] established by R. Picard and colleagues of MIT Media Lab). Such progress indicates that affective computing has indeed matured enough to have a presence and measurable impact in our lives.

In order to provide some insight on the state-of-the-art in continuous-affect-analysis-related-applications we will summarize three systems: an automatic system for human–virtual character interaction, an automatic pain detector, and a system for improving public speaking skills.

In terms of human–virtual character interaction we will look at the SEMAINE project and the Autonomous SAL system [5]. The Autonomous SAL system is unique in the current affective computing community, in that it covers the full loop of multimodal analysis, interpretation and synthesis of emotion-related and nonverbal behavior in a human-to-computer dialogue setting (see Fig. 4(a) and (b)), while at the same time being publicly available.[2,3] The system aims to engage the user in a dialogue, and create an emotional workout by paying attention to the user's nonverbal expressions, and reacting accordingly. It focuses on the 'soft skills' that humans naturally use to keep a conversation alive. The SAL characters can speak to engage the user in a simple dialogue as well as show nonverbal listener signals.

Significant progress has also been made in developing expression and affect analysis systems for behavioral science applications (see [100]). We will look at two systems, a pain detector and a public speaking skills evaluator. Creating an automatic system to detect pain in a hospital setting aims to help with efficiently monitoring patient progress. The automatic pain detector of Lucey et al. [2] (for an earlier version of the system see [1] and Fig. 1) is trained using the UNBC-McMaster database [102] that contains patients with shoulder injuries portraying real or spontaneous pain. The system detects pain at every frame, however, decision making is implemented as a post-processing step (e.g., pain intensities >10 trigger a 'pain' output). The challenge is to fine-tune the system and decide when to output a decision — the more subtle pain intensities might hinder the decision process or they might be missed.

Pfister and Robinson in [14] focus on real-time inference of nine affective states (absorbed, excited, interested, joyful, opposed, stressed, sure, thinking, and unsure) from nonverbal features of speech and using it to classify public speaking skills in terms of clear, credible, dynamic, persuasive, and competent. The classifier identifies simultaneously occurring affective states by recognizing correlations between emotions and over 6000 functional-feature combinations [164]. For multi-label affective state detection the most prominent labels describing the speech are selected. In order to apply the proposed method to real-time analysis of speech, they propose a dynamic segmentation procedure to segment the continuous audio input. For affective states classification they use pairwise-SVM and for public speaking skill assessment they utilize one binary SVM per class to derive a classwise probability. During public speaking skill assessment they detect all classes and provide a probability for each label in order to allow the participant to maximize all class probabilities (i.e., improve his speaking skills).

9. Competitions, journal Special Issues, book compilations

A major effort to advance the affective computing field is to bring together research works on various aspects of affect perception, modeling, analysis and synthesis in competitions, book compilations or journal Special Issues.

The motivation for organizing challenges is to create and provide a common benchmark data test, and to bring together various (e.g., audio and video) affect recognition communities together encouraging them to compare the relative merits of their approaches to affect recognition under well-defined and strictly comparable conditions. A second motivation is the need to advance affect recognition systems to be able to deal with naturalistic behavior in large volumes of unsegmented, non-prototypical and non-preselected data similar to what multimedia

[2] Semaine Wiki: http://semaine.opendfki.de/wiki.
[3] Semaine project website: http://www.semaine-project.eu/.

retrieval and human-machine/human-robot communication interfaces have to face in the real world.

In order to bring together and test the existing efforts for automatic affect analysis, a pioneering attempt has been that of the INTERSPEECH 2009 Emotion Challenge [190] and the INTERSPEECH 2010 Paralinguistic Challenge featuring the Affect Sub-challenge [191]. In terms of analysis in continuous input, the most noteworthy effort is the Audio/Visual Emotion Challenge and Workshop (AVEC 2011) that is the first competition event aimed at comparison of multimedia processing and machine learning methods for automatic audio, visual, and audiovisual emotion analysis [158]. Four classification problems were defined for the challenge by discretizing the activity (arousal), expectation, power, and valence dimensions as binary classification tasks, by testing at every frame whether they were above or below mean. For details and winners of the challenge the readers are referred to [158].

Affective computing has indeed matured enough to have its own IEEE Transactions and compilation books of related work papers. IEEE Transactions on Affective Computing has kick started in 2010 [192]. There are also a number of journal Special Issues that are worth mentioning for the interested readers: Affect-Based Human Behavior Understanding Special Issue [193], the Special Issue on Naturalistic Affect Resources for System Building and Evaluation [110], the Special Issue on Sensing Emotion and Affect — Facing Realism in Speech Processing [194], the Special Issue on Emotion and Mental State Recognition from Speech [195], and more recently the Special Issue on Benefits and Limitations of Continuous Representations of Emotions in Affective Computing [196].

Additionally, a number of book compilations have been put together and published in 2011 and 2012: 'Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives' [197], 'A Blueprint for Affective Computing' [198], and 'Advances in Emotion Recognition' [199]. These efforts are indeed encouraging towards bringing together the scientists from diverse backgrounds and advancing the field.

10. Concluding remarks and future directions

In line with the Special Issue on *Affect Analysis in Continuous Input*, this survey paper aimed to put the continuity aspect of affect under the spotlight by investigating the current trends in affect analysis in continuous input, and provides guidance towards possible future directions.

We will conclude this paper by providing a number of indicators for future research that may aid the researchers advance the state-of-the-art in the field, and discussing them under the titles of context, data acquisition, automatic analysis, evaluation, and system responsiveness and affect generation.

- Context. The fine-tuning of a system that will be used for affective event detection depends on the context and setting, and requires a lot of training data. Therefore, Calvo and D'Mello in [50] suggest broadening of the mental states studied, investigating dimensions versus categories for labeling, and integrating context into recognition. Context greatly affects interpretation of affect analysis in continuous input. For instance, [180] found that students who learned more tended to smile less. This suggests that the smiles that do occur may be due more to embarrassment than to a sense of achievement. People also tend to smile when experiencing frustration [200]. Such findings have implications for meaningful interpretation of sensor data, and should be carefully considered when designing automatic affect analyzers and interpreters.
- Data acquisition. When recording multiple cues it is important to keep in mind that not all recorded cues might be correlated with the displayed stimuli equally. Soleymani et al. reported that during their database recordings eye gaze was found to be more correlated with the shown content. However, the stimuli used did not induce many audio events in the participants [103]. Due to this, different detection results may be obtained with different modalities (emotion recognition results differed when using peripheral physiological signals compared to using eye gaze and EEG signals [103]).

  Single-participant (single-trial) emotion recognition requires multiple trial recordings of the same subject. This is challenging as subjects tend to get tired when recorded for long periods of time while being emotionally induced. As suggested in [103], these issues should be carefully studied prior to creating a data acquisition setup. Overall, for refining automatic analyzers, there seems to be an ever present need for (i) data with high number of participants, (ii) data with participants from diverse backgrounds in terms of culture, personality, language, age, speaking style, health state, etc., and (iii) data with richer and novel annotations by including situational context. Obtaining rich annotations is not trivial as it requires considerable amount of time and human effort. The use of Amazon's Mechanical Turk[4] has now become an alternative way among researchers to obtain statistically significant amount of annotations. Another alternative way of making the most of existing databases is to use them in a combined manner, in training and testing procedures. In other words, data from various sets can be pooled together to learn a more generic prediction model [201]. Finally, it is possible to obtain and use even larger data sets without the need for annotation. Employing unsupervised and semi-supervised learning techniques have been shown beneficial in such tasks [202]. More specifically, employing semi-supervised learning allows the exploitation of virtually infinite amounts of data available on the web (e.g., online video material). A more complex but promising alternative is to use synthesized training material, particularly for cross-corpus testing [203].
- Automatic analysis and prediction. Most of the automatic analyzers reviewed in this paper performed one-sided analysis. More specifically, most of the automatic analyzers looked only at one party (the subject) irrespective of the other party (the other person or character) they interacted with. Performing one-sided analysis may provide incomplete information regarding the affective behavior of the subject. For instance, the Duchenne smile is known to be a temporally integrated event [100]. It has also been found that during an interaction people adapt to each other, 'a shared equilibrium is formed when two people interact' [100]. Therefore, assessing reciprocity between the interacting parties (partners) is crucial for obtaining a better understanding of the interaction taking place. Pioneering studies have started to emerge in this direction (e.g., [204]). When it comes to automatic prediction of affect, confidence level of the provided predictions is also needed. Little research is found to date in this direction. Pioneering works (e.g., [3]) attempted to use labeler agreement as additional target dimension, and showed that it seems feasible to use these as system confidence for the prediction. Such confidence measures provide a measure of certainty, and are of utmost importance when integrating system decisions.
- Evaluation. When working with temporal and structured emotion data, choosing predictors that are able to optimize not only the variance of the predictor and the bias to the ground truth, but also the covariance of the prediction with respect to the ground truth, is crucial for the prediction task at hand. Emotion-specific metrics (such as SAGR) that carry valuable information regarding the emotion-specific aspects of the prediction, are also desirable [36]. Generally speaking, the performance of an automatic analyzer can be modeled and evaluated in an intrinsic and an extrinsic manner, as has been proposed for face recognition in [205]. The intrinsic performance and evaluation depend on the intrinsic components such as the data set chosen for the experiments and the machine learning algorithms and their parameters utilized for prediction.

[4] Amazon's Mechanical Turk: https://www.mturk.com/.

The extrinsic performance and evaluation instead depend on the extrinsic factors such as temporal/spatial resolution of the multimodal data, recording conditions (e.g., illumination, occlusions, noise, etc.). Future research in continuous affect detection should analyze the relevance and prospects of the aforementioned performance components, and how they could be applied to the specific problem of affect analysis in continuous input.

- System responsiveness and affect generation. An important factor for system responsiveness is the level at which the detection results should be accurate, and the level at which the detection results should be analyzed and outputted (frame-, millisecond-, second- or minute-level). These issues mainly depend on the context at hand. Despite encouraging efforts and major progress in affect analysis, highlighted in this paper, in general, dynamic social-affective capabilities of most automatic systems (e.g., virtual agents) are rather limited [206]. Overall, affect analysis and affect generation (synthesis) appear to be detached from each other even in multi-party and multi-disciplinary projects such as SEMAINE [5]. Although the overall perception and acceptability of an automated system depend on the complex interplay of these two domains, analysis and synthesis are treated as independent problems and only linked in the final stage. Investigating how to inter-relate these in earlier stages will indeed provide valuable insight into the nature of both areas that play a crucial role for the realization of affect-sensitive systems that are able to interpret multimodal and continuous input and respond appropriately.

As researchers from diverse backgrounds keep on exploring these and other future issues, we look forward to seeing automatic affect analyzers coming alive in various avenues of everyday life.

References

[1] A.B. Ashraf, S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K.M. Prkachin, P.E. Solomon, The painful face — pain expression recognition using active appearance models, Image Vision Comput. 27 (12) (2009) 1788–1796.

[2] P. Lucey, J.F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, K.M. Prkachin, Automatically detecting pain in video through facial action units, IEEE Trans. Syst. Man Cybern. Part B Cybern. 41 (3) (2011) 664–674.

[3] S. Steidl, B. Schuller, D. Seppi, A. Batliner, The hinterland of emotions: facing the open-microphone challenge, In: Proc. Int. Conf. on Affective Computing and Intelligent Interaction, IEEE, Amsterdam, The Netherlands, 2009, pp. 690–697.

[4] M. Pantic, A. Nijholt, A. Pentland, T. Huang, Human-centred intelligent human–computer interaction (hci2): how far are we from attaining it? Int. J. Auton. Adapt. Commun. Syst. (1) (2008) 168–187.

[5] M. Schröder, E. Bevacqua, R. Cowie, F. Eyben, H. Gunes, D. Heylen, M. ter Maat, G. McKeown, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. de Sevin, M. Valstar, M. Wöllmer, Building autonomous sensitive artificial listeners, IEEE Trans. Affective Comput. (2012) 1–20.

[6] M. Schröder, S. Pammi, H. Gunes, M. Pantic, M. Valstar, R. Cowie, G. McKeown, D. Heylen, M. ter Maat, F. Eyben, B. Schuller, M. Wöllmer, E. Bevacqua, C. Pelachaud, E. de Sevin, Have an emotional workout with sensitive artificial listeners! In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, p. 646.

[7] M. Schröder, E. Bevacqua, F. Eyben, H. Gunes, D. Heylen, M. Maat, S. Pammi, M. Pantic, C. Pelachaud, B. Schuller, E. Sevin, M. Valstar, M. Wöllmer, A demonstration of audiovisual sensitive artificial listeners, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, vol. 1, 2009, pp. 263–264.

[8] H. Gunes, M. Pantic, Automatic, dimensional and continuous emotion recognition, Int. J. Synth. Emotions 1 (1) (2010) 68–99.

[9] H. Gunes, M. Pantic, Automatic measurement of affect in dimensional and continuous spaces: why, what, and how? In: Proc. of Measuring Behavior, 2010, pp. 122–126.

[10] H. Gunes, B. Schuller, M. Pantic, R. Cowie, Emotion representation, analysis and synthesis in continuous space: a survey, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 827–834.

[11] D. Grandjean, D. Sander, K.R. Scherer, Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization, Conscious. Cogn. 17 (2) (2008) 484–495.

[12] P. Ekman, W. Friesen, Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues, Prentice Hall, New Jersey, 1975.

[13] S. Baron-Cohen, T. Tead, Mind Reading: The Interactive Guide to Emotion, Jessica Kingsley Publishers Ltd, 2003.

[14] T. Pfister, P. Robinson, Real-time recognition of affective states from nonverbal features of speech and its application for public speaking skill analysis, IEEE Trans. Affective Comput. 2 (2) (2011) 66–78.

[15] J.A. Russell, A circumplex model of affect, J. Pers. Soc. Psychol. 39 (1980) 1161–1178.

[16] A. Mehrabian, Pleasure–arousal–dominance: a general framework for describing and measuring individual differences in temperament, Curr. Psychol. 14 (1996) 261–292.

[17] K. Scherer, A. Schorr, T. Johnstone, Appraisal Processes in Emotion: Theory, Methods, Research, Oxford Univ. Press, Oxford/New York, 2001.

[18] C. Yu, P.M. Aoki, A. Woodruff, Detecting user engagement in everyday conversations, In: Proc. of 8th Int. Conf. on Spoken Language Processing, 2004, pp. 1329–1332.

[19] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, H.H. Chen, Music Emotion Classification: A Regression Approach, In: Proc. of IEEE Int. Conf. on Multimedia and Expo, 2007, pp. 208–211.

[20] J. Jia, S. Zhang, F. Meng, Y. Wang, L. Cai, Emotional audio-visual speech synthesis based on pad, IEEE Trans. Audio Speech Lang. Process. 19 (3) (2010) 570–582.

[21] H. Espinosa, C. Garcia, L. Pineda, Features selection for primitives estimation on emotional speech, In: Proc. of IEEE Int. Conf. on Acoustics Speech and Signal Processing, 2010, pp. 5138–5141.

[22] J.R. Fontaine, K.R. Scherer, E.B. Roesch, P. Ellsworth, The world of emotion is not two-dimensional, Psychol. Sci. 18 (2007) 1050–1057.

[23] G. McKeown, M. Valstar, R. Cowie, M. Pantic, The semaine corpus of emotionally coloured character interactions, In: Proc. of IEEE Int. Conf. Multimedia and Expo, 2010, pp. 1079–1084.

[24] R. Dietz, A. Lang, Affective agents: effects of agent affect on arousal, attention, liking and learning, In: Proc. of Cognitive Technology, 1999.

[25] C. Kaernbach, On dimensions in emotion psychology, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 792–796.

[26] N.H. Frijda, The Emotions, Cambridge Univ. Press, 1986.

[27] D. Sander, D. Grandjean, K.R. Scherer, A systems approach to appraisal mechanisms in emotion, Neural Networks 18 (4) (2005) 317–352.

[28] M. Mortillaro, B. Meuleman, K.R. Scherer, Advocating a componential appraisal model to guide emotion recognition, J. Synth. Emotions 3 (1) (2012) 18–32.

[29] A. Vinciarelli, M. Pantic, H. Bourlard, Social signal processing: survey of an emerging domain, Image Vision Comput. 27 (12) (2009) 1743–1759.

[30] A. Tarasov, S.J. Delany, Benchmarking classification models for emotion recognition in natural speech: a multi-corporal study, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 841–846.

[31] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge, J. Speech Commun. 53 (9–10) (2011) 1062–1087.

[32] G. Chanel, J.J.M. Kierkels, M. Soleymani, T. Pun, Short-term emotion assessment in a recall paradigm, Int. J. Hum. Comput. Stud. 67 (8) (2009) 607–627.

[33] G. Berntson, J. Bigger, D. Eckberg, P. Grossman, P. Kaufmann, M. Malik, H. Nagaraja, S. Porges, J. Saul, P. Stone, M. VanderMolen, Heart rate variability:origins, methods, and interpretive caveats, Psychophysiology 34 (6) (1997) 623–648.

[34] L. Salahuddin, J. Cho, M.G. Jeong, D. Kim, Ultra short term analysis of heart rate variability for monitoring mental stress in mobile settings, In: Proc. of the IEEE Int. Conf. of Eng. Med. Biol. Soc., 2007, pp. 39–48.

[35] A. Nakasone, H. Prendinger, M. Ishizuka, Emotion recognition from electromyography and skin conductance, In: Proc. of the 5th Int. Workshop on Biosignal Interpretation, 2005, pp. 219–222.

[36] M. Nicolaou, H. Gunes, M. Pantic, Continuous prediction of spontaneous affect from multiple cues and modalities in valence–arousal space, IEEE Trans. Affective Comput. 2 (2) (2011) 92–105.

[37] S. Petridis, H. Gunes, S. Kaltwang, M. Pantic, Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities, In: Proc. of ACM Int. Conf. on Multimodal Interfaces, 2009, pp. 23–30.

[38] B. Schuller, L. Devillers, Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm, In: Proc. of INTERSPEECH, 2010, pp. 801–804.

[39] H. Gunes, M. Piccardi, M. Pantic, Affective computing: focus on emotion expression, synthesis, and recognition, In: Ch. From the Lab to the Real World: Affect Recognition using Multiple Cues and Modalities, I-Tech Education and Publishing, 2008, pp. 185–218.

[40] Z. Zeng, M. Pantic, G. Roisman, T. Huang, A survey of affect recognition methods: audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2009) 39–58.

[41] G. Chanel, K. Ansari-Asl, T. Pun, Valence–arousal evaluation using physiological signals in an emotion recall paradigm, In: Proc. of IEEE Int. Conf. on Systems, Man and Cybernetics, 2007, pp. 2662–2667.

[42] A. Haag, S. Goronzy, P. Schaich, J. Williams, Emotion recognition using bio-sensors: first steps towards an automatic system, In: LNCS, 3068, 2004, pp. 36–48.

[43] T. Pun, T. Alecu, G. Chanel, J. Kronegg, S. Voloshynovskiy, Brain–computer interaction research at the computer vision and multimedia laboratory, University of Geneva, IEEE Trans. Neural Syst. Rehabil. Eng. 14 (2) (2006) 210–213.

[44] R. Davidson, N. Fox, Asymmetrical brain activity discriminates between positive and negative affective stimuli in human infants, Science 218 (1982) 1235–1237.

[45] P. Tsiamyrtzis, J. Dowdall, D. Shastri, I. Pavlidis, M. Frank, P. Ekman, Imaging facial physiology for the detection of deceit, Int. J. Comput. Vision 71 (2) (2007) 197–214.

[46] M. Kusserow, O. Amft, G. Troster, Bodyant: miniature wireless sensors for naturalistic monitoring of daily activity, In: Proc. of Int. Conf. on Body Area Networks, 2009, pp. 1–8.

[47] Emotiv's epoc, http://www.emotiv.com/.

[48] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, M. Schröder, Feeltrace: an instrument for recording perceived emotion in real time, In: Proc. of ISCA Workshop on Speech and Emotion, 2000, pp. 19–24.

[49] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J.G. Taylor, Emotion recognition in human–computer interaction, IEEE Signal Process. Mag. 18 (1) (2001) 33–80.

[50] R. Calvo, S. D'Mello, Affect detection: an interdisciplinary review of models, methods, and their applications, IEEE Trans. Affective Comput. 1 (1) (2010) 18–37.
[51] G.L. Huttar, Relations between prosodic variables and emotions in normal american english utterances, J. Speech Hear. Res. 11 (1968) 481–487.
[52] J. Davitz, The communication of emotional meaning, In: Ch. Auditory Correlates of Vocal Expression of Emotional Feeling, McGraw-Hill, 1964, pp. 101–112.
[53] K.R. Scherer, J.S. Oshinsky, Cue utilization in emotion attribution from auditory stimuli, Motiv. Emot. 1 (1977) 331–346.
[54] J. Trouvain, W.J. Barry, The prosody of excitement in horse race commentaries, In: Proc. of ISCA Workshop Speech Emotion, 2000, pp. 86–91.
[55] M. Schröder, Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis, Ph.D. dissertation, Univ. of Saarland, Germany, 2003.
[56] M. Schröder, D. Heylen, I. Poggi, Perception of non-verbal emotional listener feedback, In: in: R. Hoffmann, H. Mixdorff (Eds.), Speech Prosody, 2006, pp. 1–4.
[57] M. Wöllmer, B. Schuller, F. Eyben, G. Rigoll, Combining long short-term memory and dynamic bayesian networks for incremental emotion-sensitive artificial listening, IEEE J. Sel. Top. Sign. Proces. 4 (5) (2010) 867–881.
[58] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, H. Konosu, Being bored? Recognising natural interest by extensive audiovisual integration for real-life application, Image Vision Comput. J. 27 (12) (2009) 1760–1774.
[59] A. Batliner, S. Steidl, F. Eyben, B. Schuller, On laughter and speech laugh, based on observations of child-robot interaction, In: in: J. Trouvain, N. Campbell (Eds.), The Phonetics of Laughing, Saarland University Press, Saarbrücken, 2011.
[60] I. Becker, V. Aharonson, Last but definitely not least: on the role of the last sentence in automatic polarity-classification, In: Proc. of the ACL 2010 Conference, 2010, pp. 331–335.
[61] K. Matsumoto, F. Ren, Estimation of word emotions based on part of speech and positional information, Comput. Hum. Behav. 27 (5) (2011) 1553–1564.
[62] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, V. Subrahmanian, Sentiment analysis: adjectives and adverbs are better than adjectives alone, In: Proc. Int. Conf. on Weblogs and Social Media, 2007, pp. 1–7.
[63] V. Subrahmanian, D. Reforgiato, AVA: adjective–verb–adverb combinations for sentiment analysis, Intell. Syst. 23 (4) (2008) 43–50.
[64] M. Pantic, M. Bartlett, Machine analysis of facial expressions, In: in: K. Delac, M. Grgic (Eds.), Face Recognition, I-Tech Education and Publishing, Vienna, Austria, 2007, pp. 377–416.
[65] N. Dael, M. Mortillaro, K.R. Scherer, The body action and posture coding system (bap): Development and reliability, J. Nonverbal Behav. 36 (2) (2012) 97–121 (Springer).
[66] H. Gunes, C. Shan, S. Chen, Y. Tian, Advances in emotion recognition, In: Ch. Bodily Expression for Automatic Affect Recognition, Wiley-Blackwell, 2012, pp. 1–35.
[67] C. Darwin, The Expression of the Emotions in Man and Animals, John Murray, London, 1872.
[68] H. Wallbott, Bodily expression of emotion, Eur. J. Soc. Psychol. 28 (1998) 879–896.
[69] F. Pollick, H. Paterson, A. Bruderlin, A. Sanford, Perceiving affect from arm movement, Cognition 82 (2001) 51–61.
[70] M. Coulson, Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence, Nonverbal Behav. 28 (2) (2004) 117–139.
[71] A. Kleinsmith, N. Bianchi-Berthouze, Recognizing affective dimensions from body posture, In: Proc. of the Int. Conf. on Affective Computing and Intelligent Interaction, 2007, pp. 48–58.
[72] H. Gunes, M. Piccardi, Automatic temporal segment detection and affect recognition from face and body display, IEEE Trans. Syst. Man Cybern. B Cybern. 39 (1) (2009) 64–84.
[73] J. Cohn, L. Reed, T. Moriyama, J. Xiao, K. Schmidt, Z. Ambadar, Multimodal coordination of facial action, head rotation, and eye motion during spontaneous smiles, In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2004, pp. 129–135.
[74] R. Cowie, H. Gunes, G. McKeown, L. Vaclau-Schneider, J. Armstrong, E. Douglas-Cowie, The emotional and communicative significance of head nods and shakes in a naturalistic database, In: Proc. of LREC Int. Workshop on Emotion, 2010, pp. 42–46.
[75] H. Gunes, M. Pantic, Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners, In: Proc. of Int. Conf. on Intelligent Virtual Agents, 2010, pp. 371–377.
[76] A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: a survey, IEEE Trans. Affective Comput. 99 (PP) (2012) 1–20.
[77] M. Chen, J.A. Bargh, Consequences of automatic evaluation: immediate behavioral predispositions to approach or avoid the stimulus, Pers. Soc. Psychol. Bull. 25 (2) (1999) 215–224.
[78] J. Forster, F. Strack, Influence of overt head movements on memory for valenced words: a case of conceptual–motor compatibility, J. Pers. Soc. Psychol. 71 (1996) 421–430.
[79] C. Carver, Pleasure as a sign you can attend to something else: placing positive feelings within a general model of affect, Cogn. Emotion 17 (2) (2003) 241–261.
[80] C. Hillman, K. Rosengren, D. Smith, Emotion and motivated behavior: postural adjustments to affective picture viewing, Biol. Psychol. 66 (2004) 51–62.
[81] D. Janssen, W.I. Schllhorn, J. Lubienetzki, K. Fling, H. Kokenge, K. Davids, Recognition of emotions in gait patterns by means of artificial neural nets, J. Nonverbal Behav. 32 (2008) 79–92.
[82] M. Karg, K. Khnlenz, M. Buss, Recognition of affect based on gait patterns, IEEE Trans. Syst. Man Cybern. B Cybern. 40 (2010) 1050–1061.
[83] M. Inderbitzin, A. Väljamäe, J.M.B. Calvo, Expression of emotional states during locomotion based on canonical parameters, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 809–814.
[84] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, S. Narayanan, Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling, In: Proc. of INTERSPEECH, 2010, pp. 2362–2365.
[85] A. Kleinsmith, P.R. De Silva, N. Bianchi-Berthouze, Recognizing emotion from postures: cross–cultural differences in user modeling, In: Proc. of the Conf. on User Modeling, 2005, pp. 50–59.
[86] A. Metallinou, A. Katsamanis, Y. Wang, S. Narayanan, Tracking changes in continuous emotion states using body language and prosodic cues, In: Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2011, pp. 2288–2291.
[87] H.K. Meeren, C.C. Van Heijnsbergen, B. De Gelder, Rapid perceptual integration of facial expression and emotional body language, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 16518–16523.
[88] J. Van den Stock, R. Righart, B. De Gelder, Body expressions influence recognition of emotions in the face and voice, Emotion 7 (3) (2007) 487–494.
[89] W. Labov, Language in use, In: Ch. Field Methods of the Project in Linguistic Change and Variation, Prentice-Hall, 1984, pp. 28–53.
[90] H. Brugman, A. Russel, Annotating multi-media/multi-modal resources with ELAN, In: Proc. of Int. Conf. on Language Resources and Evaluation, 2004, pp. 2065–2068.
[91] M. Kipp, Anvil — a generic annotation tool for multimodal dialogue, In: Proc. of the 7th European Conference on Speech Communication and Technology, 2001, pp. 1367–1370.
[92] R. Cowie, G. McKeown, E. Douglas-Cowie, Tracing emotion: an overview, J. Synth. Emotions 3 (1) (2012) 1–17.
[93] D. Kulic, E.A. Croft, Affective state estimation for human–robot interaction, IEEE Trans. Robot. 23 (5) (2007) 991–1000.
[94] B. Schuller, S. Hantke, F. Weninger, W. Han, Z. Zhang, S. Narayanan, Automatic recognition of emotion evoked by general sound events, In: Proceedings 37th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012, 2012, pp. 341–344.
[95] P. Lang, The Cognitive Psychophysiology of Emotion: Anxiety and the Anxiety Disorders, Lawrence Erlbaum, Hillside, NJ, 1985.
[96] G. Laurans, P. Desmet, P. Hekkert, The emotion slider: a self-report device for the continuous measurement of emotion, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction Workshops, 2009, pp. 1–6.
[97] J. Kessens, M. Neerincx, R. Looije, M. Kroes, G. Bloothooft, Perception of synthetic emotion expressions in speech: categorical and dimensional annotations, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction Workshops, 2009, pp. 1–5.
[98] M. Grimm, K. Kroschel, Evaluation of natural emotions using self assessment manikins, In: Proc. ASRU, 2005, pp. 381–385.
[99] M. Grimm, K. Kroschel, S. Narayanan, The Vera am Mittag German audio-visual emotional speech database, In: Proc. of IEEE Int. Conf. on Multimedia and Expo, 2008, pp. 865–868.
[100] J. Cohn, Advances in behavioral science using automated facial image analysis and synthesis, IEEE Signal Process. Mag. (2010) 128–133.
[101] J. Healey, Recording affect in the field: towards methods and metrics for improving ground truth labels, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, 2011, pp. 107–116.
[102] P. Lucey, J.F. Cohn, K.M. Prkachin, P.E. Solomon, I. Matthews, Painful data: the UNBC-McMaster shoulder pain expression archive database, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 57–64.
[103] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multi-modal database for affect recognition and implicit tagging, IEEE Trans. Affective Comput. 3 (1) (2012) 42–55.
[104] S. Koelstra, C. Muehl, I. Patras, EEG analysis for implicit tagging of video data, In: Proc. of the Affective Brain–Computer Interfaces Workshop, 2009, pp. 27–32.
[105] S. Koelstra, C. Mühl, M. Soleymani, A. Yazdani, J.-S. Lee, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: a database for emotion analysis using physiological signals, IEEE Trans. Affective Comput. 3 (1) (2012) 18–31.
[106] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schröder, The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent, IEEE Trans. Affective Comput. 3 (1) (2012) 5–17.
[107] I. Sneddon, M. McRorie, G. McKeown, J. Hanratty, The Belfast induced natural emotion database, IEEE Trans. Affective Comput. 3 (1) (2012) 32–41.
[108] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, S. Narayanan, The USC CreativeIT database: a multimodal database of theatrical improvisation, In: Proc. of LREC Workshop on Multimodal Corpora, 2010.
[109] X. Sun, J. Lichtenauer, M. Valstar, A. Nijholt, M. Pantic, A multimodal database for mimicry analysis, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, 2011, pp. 367–376.
[110] B. Schuller, E. Douglas-Cowie, A. Batliner, Guest editorial: special section on naturalistic affect resources for system building and evaluation, IEEE Trans. Affective Comput. 3 (1) (2012) 3–4.
[111] H.P. Espinosa, L.V.P.C.A.R. Garcia, Bilingual acoustic feature selection for emotion estimation using a 3D continuous model, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 786–791.
[112] M.C. Sezgin, B. G-nsel, G.K. Kurt, A novel perceptual feature set for audio emotion recognition, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 780–785.

[113] J. Snel, C. Cullen, Obtaining speech assets for judgement analysis on low-pass filtered emotional speech, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 835–840.

[114] M. Nicolaou, H. Gunes, M. Pantic, Output-associative RVM regression for dimensional and continuous emotion prediction, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 16–23.

[115] G. Caridakis, K. Karpouzis, S. Kollias, User and context adaptive neural networks for emotion recognition, Neurocomputer 71 (13–15) (2008) 2553–2562.

[116] I. Kanluan, M. Grimm, K. Kroschel, Audio-visual emotion recognition using an emotion recognition space concept, Proc. of the 16th European Signal Processing Conference, 2008.

[117] K. Forbes-Riley, D. Litman, Predicting emotion in spoken dialogue from multiple knowledge sources, In: Proc. of Human Language Technology Conf. North Am. Chapter of the Assoc. Computational Linguistics, 2004, pp. 201–208.

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

[119] J. Kim, Robust speech recognition and understanding, In: Ch. Bimodal Emotion Recognition using Speech and Physiological Changes, I-Tech Education and Publishing, 2007, pp. 265–280.

[120] M. Nicolaou, H. Gunes, M. Pantic, Audio-visual classification and fusion of spontaneous affective data in likelihood space, In: Proc. of IEEE Int. Conf. on Pattern Recognition, 2010, pp. 3695–3699.

[121] C. Liu, P. Rani, N. Sarkar, An empirical study of machine learning techniques for affect recognition in human-robot interaction, In: Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, 2005, pp. 2662–2667.

[122] R. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state, IEEE Trans. Pattern Anal. Mach. Intell. 23 (10) (2001) 1175–1191.

[123] Y. Gu, S.-L. Tan, K.-J. Wong, M.-H.R. Ho, L. Qu, Emotion-aware technologies for consumer electronics, In: Proc. of IEEE Int. Symp. on Consumer Electronics, 2008, pp. 1–4.

[124] R. Khosrowabadi, H.C. Quek, A. Wahab, K.K. Ang, EEG-based emotion recognition using self-organizing map for boundary detection, In: Proc. of Int. Conf. on Pattern Recognition, 2010, pp. 4242–4245.

[125] C. Frantzidis, C. Bratsas, M. Klados, E. Konstantinidis, C. Lithari, A. Vivas, C. Papadelis, E. Kaldoudi, C. Pappas, P. Bamidis, On the classification of emotional biosignals evoked while viewing affective pictures: an integrated data-mining-based approach for healthcare applications, IEEE Trans. Inf. Technol. Biomed. 14 (2) (2010) 309–318.

[126] D. Heger, F. Putze, T. Schultz, Online recognition of facial actions for natural EEG-based BCI applications, In: Proc. of ACII 2011 Affective Brain-Computer Interfaces Workshop, 2011, pp. 436–446.

[127] B.Z.A.S.D.D.H. Christian Mhl, Anton Nijholt, Nijholt, affective brain–computer interfaces, In: Proc. of ACII 2011 Affective Brain–Computer Interfaces Workshop, 2011, p. 435.

[128] T.O. Zander, C. Kothe, Towards passive brain–computer interfaces: applying brain–computer interface technology to human-machine systems in general, J. Neural Eng. 8 (2011) 1–5.

[129] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, A. Wendemuth, Acoustic emotion recognition: a benchmark comparison of performances, In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, 2009, pp. 552–557.

[130] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, Abandoning emotion classes — towards continuous emotion recognition with modelling of long-range dependencies, In: Proc. of 9th Interspeech Conf., 2008, pp. 597–600.

[131] M. Grimm, K. Kroschel, Emotion estimation in speech using a 3D emotion space concept, In: Proc. of IEEE Automatic Speech Recognition and Understanding Workshop, 2005, pp. 381–385.

[132] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, L. Lowry, M. McRorie, L. Jean-Claude Martin, J.-C. Devillers, A. Abrilian, S. Batliner, A. Noam, K. Karpouzis, The humaine database: addressing the needs of the affective computing community, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, 2007, pp. 488–500.

[133] A. Reyes, P. Rosso, Linking humour to blogs analysis: affective traits in posts, In: Proc. Int. Workshop on Opinion Mining and Sentiment Analysis, 2009, pp. 205–212.

[134] C. Strapparava, R. Mihalcea, Annotating and identifying emotions in text, In: in: G. Armano, M. de Gemmis, G. Semeraro, E. Vargiu (Eds.), Intelligent Information Access, Vol. 301 of Studies in Computational Intelligence, Springer, Berlin, Heidelberg, 2010, pp. 21–38.

[135] A. Balahur, J.M. Hermida, A. Montoyo, Detecting emotions in social affective situations using the EmotiNet knowledge base, In: Proc. International Symposium on Neural Networks, IEEE, vol. 3, Guilin, China, 2011, pp. 611–620.

[136] B. Schuller, Recognizing affect from linguistic information in 3D continuous space, IEEE Trans. Affect. Comput. 2 (4) (2011) 192–205.

[137] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, N. Amir, Whodunnit — searching for the most important feature types signalling emotion-related user states in speech, Comp. Speech Lang. 25 (1) (2011) 4–28.

[138] F. Eyben, M. Wöllmer, M. Valstar, H. Gunes, B. Schuller, M. Pantic, String-based audiovisual fusion of behavioural events for the assessment of dimensional affect, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 322–329.

[139] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, S. Steidl, Emotion recognition using imperfect speech recognition, In: Proc. INTERSPEECH, Makuhari, Japan, 2010, pp. 478–481.

[140] B. Schuller, T. Knaup, Learning and knowledge-based sentiment analysis in movie review key excerpts, In: in: A. Esposito, A. Esposito, R. Martone, V. Müller, G. Scarpetta (Eds.), Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues, vol. 6456/2010, Springer Lecture Notes on Computer Science (LNCS), 2010, pp. 448–472.

[141] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Senticspace: visualizing opinions and sentiments in a multi-dimensional vector space, In: in: R. Setchi, I. Jordanov, R. Howlett, L. Jain (Eds.), Knowledge-Based and Intelligent Information and Engineering Systems, Vol. 6279 of Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2010, pp. 385–393.

[142] R. Poppe, Vision-based human motion analysis: an overview, Comput. Vision and Image Understanding 108 (1–2) (2007) 4–18.

[143] R. Poppe, A survey on vision-based human action recognition, Image Vision Comput. 28 (6) (2010) 976–990.

[144] D. McDuff, R. El Kaliouby, K. Kassam, R. Picard, Affect valence inference from facial action unit spectrograms, In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2010, pp. 17–24.

[145] D. Glowinski, A. Camurri, G. Volpe, N. Dael, K. Scherer, Technique for automatic emotion recognition by body gesture analysis, In: Proc. of Computer Vision and Pattern Recognition Workshops, 2008, pp. 1–6.

[146] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, S. Kollias, Emotion recognition through facial expression analysis based on a neurofuzzy method, J. Neural Networks 18 (2005) 423–435.

[147] S. Arifin, P. Cheung, Affective level video segmentation by utilizing the pleasure–arousal–dominance information, IEEE Trans. Multimedia 10 (7) (2008) 1325–1341.

[148] M. Nicolaou, H. Gunes, M. Pantic, Automatic segmentation of spontaneous data using dimensional labels from multiple coders, In: Proc. of LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, 2010, pp. 43–48.

[149] M. Kipp, J.-C. Martin, Gesture and emotion: can basic gestural form features discriminate emotions? In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction Workshops, 2009, pp. 1–8.

[150] Y. Yoshitomi, S.I. Kim, T. Kawano, T. Kitazoe, Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face, In: IEEE Int. Workshop on Robot and Human Interactive Communication, 2000, pp. 178–183.

[151] M.M. Khan, R.D. Ward, M. Ingleby, Infrared thermal sensing of positive and negative affective states, In: Proc. of the IEEE Int. Conf. on Robotics, Automation and Mechatronics, 2006, pp. 1–6.

[152] B. Nhan, T. Chau, Classifying affective states using thermal infrared imaging of the human face, IEEE Trans. Biomed. Eng. 57 (4) (2010) 979–987.

[153] A. Merla, G. Romani, Thermal signatures of emotional arousal: a functional infrared imaging study, In: Proc. of Int. Conf. of the IEEE Engineering in Medicine and Biology Society, 2007, pp. 247–249.

[154] S. Gilroy, M. Cavazza, M. Niiranen, E. Andre, T. Vogt, J. Urbain, M. Benayoun, H. Seichter, M. Billinghurst, Pad-based multimodal affective fusion, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction Workshops, 2009, pp. 1–8.

[155] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, K. Karpouzis, Modeling naturalistic affective states via facial and vocal expressions recognition, In: Proc. of ACM Int. Conf. on Multimodal Interfaces, 2006, pp. 146–154.

[156] A.M. Oliveira, M.P. Teixeira, I.B. Fonseca, M. Oliveira, Joint model-parameter validation of self-estimates of valence and arousal: probing a differential-weighting model of affective intensity, In: Proc. of the 22nd Annual Meeting of the Int. Society for Psychophysics, 2006, pp. 245–250.

[157] N. Alvarado, Arousal and valence in the direct scaling of emotional response to film clips, Motiv. Emot. 21 (4) (1997) 323–348.

[158] B. Schuller, M. Valstar, R. Cowie, M. Pantic, Avec 2011 — the first audio/visual emotion challenge and workshop — an introduction, In: Proc. of 1st Int'l. Audio/-Visual Emotion Challenge and Workshop, 2011, pp. 415–424.

[159] M. Nicolaou, H. Gunes, M. Pantic, Designing frameworks for automatic affect prediction and classification in dimensional space, In: Proc. of IEEE CVPR Workshop on Gesture Recognition, 2011, pp. 20–26.

[160] J. Kim, E. Andre, Emotion recognition based on physiological changes in music listening, IEEE Trans. Pattern Anal. Mach. Intell. 30 (12) (2008) 2067–2083.

[161] M. Nicolaou, H. Gunes, M. Pantic, A multi-layer hybrid framework for dimensional emotion classification, In: Proc. of ACM Multimedia, 2011, pp. 933–936.

[162] S. Bermejo, J. Cabestany, Oriented principal component analysis for large margin classifiers, Neural Networks 14 (10) (2001) 1447–1461.

[163] M. Schröder, The SEMAINE API: Towards a standards-based framework for building emotion-oriented systems, Adv. Hum. Mach. Interact. 2010 (2010) 1–21.

[164] F. Eyben, M. Wöllmer, B. Schuller, OpenSMILE — the munich versatile and fast open-source audio feature extractor, In: Proc. of ACM Multimedia, 2010, pp. 1459–1462.

[165] T. Taleb, D. Bottazzi, N. Nasser, A novel middleware solution to improve ubiquitous healthcare systems aided by affective information, IEEE Trans. Inf. Technol. Biomed. 14 (2) (2010) 335–349.

[166] G. Littlewort, J. Whitehill, T. Wu, I.R. Fasel, M.G. Frank, J.R. Movellan, M.S. Bartlett, The computer expression recognition toolbox (cert), In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 298–305.

[167] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, K. Scherer, Towards a minimal representation of affective gestures, IEEE Trans. Affective Comput. 2 (2) (2011) 106–118.

[168] Humaine: http://emotion–research.net.

[169] T. Huang, M. Hasegawa-Johnson, S. Chu, Z. Zeng, H. Tang, Sensitive talking heads, IEEE Signal Process. Mag. 26 (4) (2009) 67–72.
[170] X. Shen, X. Fu, Y. Xuan, Do different emotional valences have same effects on spatial attention? In: Proc. of Int. Conf. on Natural Computation, vol. 4, 2010, pp. 1989–1993.
[171] K.C. Wassermann, K. Eng, P.F.M.J. Verschure, Live soundscape composition based on synthetic emotions, IEEE MultiMedia 10 (4) (2003) 82–90.
[172] A. Beck, L. Canamero, K. Bard, Towards an affect space for robots to display emotional body language, In: Proc. of IEEE Int. Symp. in Robot and Human Interactive Communication, 2010, pp. 464–469.
[173] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P.W. McOwan, A. Paiva, Automatic analysis of affective postures and body motion to detect engagement with a game companion, In: Proc. of ACM/IEEE Int. Conf. on Human-Robot Interaction, 2011, pp. 305–311.
[174] M. Karg, M. Schwimmbeck, K. Kühnlenz, M. Buss, Towards mapping emotive gait patterns from human to robot, In: Proc. of IEEE Int. Symp. in Robot and Human Interactive Communication, 2010, pp. 258–263.
[175] M. Mihelj, D. Novak, M. Munih, Emotion-aware system for upper extremity rehabilitation, In: Proc. of Int. Conf. on Virtual Rehabilitation, 2009, pp. 160–165.
[176] T.-C. Tsai, J.-J. Chen, W.-C. Lo, Design and implementation of mobile personal emotion monitoring system, In: Proc. of Int. Conf. on Mobile Data Management: Systems, Services and Middleware, 2009, pp. 430–435.
[177] B. Grundlehner, L. Brown, J. Penders, B. Gyselinckx, The design and analysis of a real-time, continuous arousal monitor, In: Proc. of Int. Workshop on Wearable and Implantable Body Sensor Networks, 2009, pp. 156–161.
[178] R. Picard, Future affective technology for autism and emotion communication, Philos. Trans. R. Soc. Lond. B Biol. Sci. 364 (1535) (2009) 3575–3584.
[179] G. Littlewort, M.S. Bartlett, L.P. Salamanca, J. Reilly, Automated measurement of children's facial expressions during problem solving tasks, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 30–35.
[180] J. Whitehill, Z. Serpell, A. Foster, Y. Lin, B. Pearson, M.S. Barlett, J.R. Movellan, Toward an optimal affect-sensitive instructional system of cognitive skills, In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition Workshop on Human Communicative Behavior, 2011, pp. 20–25.
[181] U. Faghihi, P. Fournier-Viger, R. Nkambou, P. Poirier, A. Mayers, How emotional mechanism helps episodic learning in a cognitive agent, In: Proc. of IEEE Symp. on Intelligent Agents, 2009, pp. 23–30.
[182] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, N. Nguyen-Thien, Emotion on the road — necessity, acceptance, and feasibility of affective computing in the car, Adv. Hum. Mach. Interact. 2010 (2010) 1–17.
[183] K. Sun, J. Yu, Y. Huang, X. Hu, An improved valence–arousal emotion space for video affective content representation and recognition, In: Proc. of IEEE Int. Conf. on Multimedia and Expo, 2009, pp. 566–569.
[184] J. Kierkels, M. Soleymani, T. Pun, Queries and tags in affect-based multimedia retrieval, In: Proc. of IEEE Int. Conf. on Multimedia and Expo, 2009, pp. 1436–1439.
[185] M. Soleymani, J. Davis, T. Pun, A collaborative personalized affective video retrieval system, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction Workshops, 2009, pp. 1–2.
[186] M. Soleymani, S. Koelstra, I. Patras, T. Pun, Continuous emotion detection in response to music videos, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 803–808.
[187] M. Rehm, M. Wissner, Gamble—a multiuser game with an embodied conversational agent, In: Lecture Notes in Computer Science, vol. 3711, 2005, pp. 180–191.
[188] A. Kleinsmith, N. Bianchi-Berthouze, A. Steed, Automatic recognition of non-acted affective postures, IEEE Trans. Syst. Man Cybern. B Cybern. 41 (2011) 1027–1038.
[189] Affectiva's homepage: http://www.affectiva.com/.
[190] B. Schuller, S. Steidl, A. Batliner, The interspeech 2009 emotion challenge, In: Proceedings INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, 2009, pp. 312–315.
[191] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. M-Iler, S. Narayanan, The interspeech 2010 paralinguistic challenge, In: Proc. of INTERSPEECH, 2010, pp. 2794–2797.
[192] J. Gratch, Editorial, IEEE Trans. Affective Comput. 1 (1) (2010) 1–10.
[193] A.A. Salah, T. Gevers, A. Vinciarelli, Introduction to the affect-based human behavior understanding special issue, IEEE Trans. Affective Comput. 2 (2) (2011) 64–65.
[194] B. Schuller, S. Steidl, A. Batliner, Introduction to the special issue on sensing emotion and affect — facing realism in speech processing, Speech Commun. 53 (9/10) (2011) 1059–1061.
[195] J. Epps, R. Cowie, S. Narayanan, B. Schuller, J. Tao, Editorial emotion and mental state recognition from speech, EURASIP J. Adv. Signal Process. 2012 (15) (2012) 1–2.
[196] E. Hudlicka, H. Gunes, Benefits and limitations of continuous representations of emotions in affective computing: introduction to the special issue, J. Synth. Emotions 3 (1) (2012) (i–vi).
[197] D. Gokcay, G. Yildirim, Affective Computing and Interaction: Psychological, Cognitive and Neuroscientific Perspectives, IGI Global, 2011.
[198] K.R. Scherer, T. B-nziger, E. Roesch, A Blueprint for Affective Computing, Oxford University Press, 2011.
[199] A. Konar, A. Chakraborty, Advances in Emotion Recognition, Wiley-Blackwell, 2012.
[200] M.E. Hoque, R.W. Picard, Acted vs. natural frustration and delight: many people smile in natural frustration, In: Proc. of IEEE Int. Conf. on Automatic Face and Gesture Recognition, 2011, pp. 354–359.
[201] B. Schuller, Z. Zhang, F. Weninger, G. Rigoll, Using multiple databases for training in emotion recognition: to unite or to vote? In: Proc. INTERSPEECH, 2011, pp. 1553–1556.
[202] Z. Zhang, F. Weninger, M. Wöllmer, B. Schuller, Unsupervised learning in cross-corpus acoustic emotion recognition, In: Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding, 2011, pp. 523–528.
[203] B. Schuller, F. Burkhardt, Learning with synthesized speech for automatic emotion recognition, In: Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2010, pp. 5150–5153.
[204] X. Sun, A. Nijholt, K.P. Truong, M. Pantic, Automatic understanding of affective and social signals by multimodal mimicry recognition, In: Proc. of Int. Conf. on Affective Computing and Intelligent Interaction, 2011, pp. 289–296.
[205] P. Wang, Q. Ji, Performance modeling and prediction of face recognition systems, In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1566–1573.
[206] D. Heylen, M. Theune, R. op den Akker, A. Nijholt, Social agents: the first generations, In: Proc. of Affective Computing and Intelligent Interaction Workshops, 2009, pp. 1–7.