A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions

Zhihong Zeng, *Member*, *IEEE Computer Society*, Maja Pantic, *Senior Member*, *IEEE*, Glenn I. Roisman, and Thomas S. Huang, *Fellow*, *IEEE*

Abstract—Automated analysis of human affective behavior has attracted increasing attention from researchers in psychology, computer science, linguistics, neuroscience, and related disciplines. However, the existing methods typically handle only deliberately displayed and exaggerated expressions of prototypical emotions, despite the fact that deliberate behavior differs in visual appearance, audio profile, and timing from spontaneously occurring behavior. To address this problem, efforts to develop algorithms that can process naturally occurring human affective behavior have recently emerged. Moreover, an increasing number of efforts are reported toward multimodal fusion for human affect analysis, including audiovisual fusion, linguistic and paralinguistic fusion, and multicue visual fusion based on facial expressions, head movements, and body gestures. This paper introduces and surveys these recent advances. We first discuss human emotion perception from a psychological perspective. Next, we examine available approaches for solving the problem of machine understanding of human affective behavior and discuss important issues like the collection and availability of training and test data. We finally outline some of the scientific and engineering challenges to advancing human affect sensing technology.

Index Terms—Evaluation/methodology, human-centered computing, affective computing, introductory, survey.

1 Introduction

A widely accepted prediction is that computing will move to the background, weaving itself into the fabric of our everyday living spaces and projecting the human user into the foreground. Consequently, the future "ubiquitous computing" environments will need to have human-centered designs instead of computer-centered designs [26], [31], [100], [107], [109]. Current human-computer interaction (HCI) designs, however, usually involve traditional interface devices such as the keyboard and mouse and are constructed to emphasize the transmission of explicit messages while ignoring implicit information about the user, such as changes in the affective state. Yet, a change in the user's affective state is a fundamental component of human-human communication. Some affective states motivate human actions, and others enrich the meaning of human communication. Consequently, the traditional HCI, which ignores the user's affective states, filters out a large portion of the information available in the interaction process. As a result, such interactions are frequently perceived as cold, incompetent, and socially inept. The human computing paradigm suggests that user interfaces of the future need to be anticipatory and human centered, built for humans, and based on naturally occurring multimodal human communication [100], [109]. Specifically, human-centered interfaces must have the ability to detect subtleties of and changes in the user's behavior, especially his/her affective behavior, and to initiate interactions based on this information rather than simply responding to the user's commands.

Examples of affect-sensitive multimodal HCI systems include the following:

1. the system of Lisetti and Nasoz [85], which combines facial expression and physiological signals to recognize the user's emotions, like fear and anger, and then to adapt an animated interface agent to mirror the user's emotion,
2. the multimodal system of Duric et al. [39], which applies a model of embodied cognition that can be seen as a detailed mapping between the user's affective states and the types of interface adaptations,
3. the proactive HCI tool of Maat and Pantic [89], which is capable of learning and analyzing the user's context-dependent behavioral patterns from multisensory data and of adapting the interaction accordingly,
4. the automated Learning Companion of Kapoor et al. [72], which combines information from cameras, a sensing chair, and mouse, wireless skin sensor, and task state to detect frustration in order to predict when the user needs help, and

- *Z. Zeng and T.S. Huang are with the Beckman Institute, University of Illinois at Urbana-Champaign, 405 N. Mathews Ave., Urbana, 61801. E-mail: {zhzeng, huang}@ifp.uiuc.edu.*
- *M. Pantic is with the Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, United Kingdom, and with the Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, The Netherlands. E-mail: m.pantic@imperial.ac.uk.*
- *G.I. Roisman is with the Psychology Department, University of Illinois at Urbana-Champaign, 603 East Daniel St., Champaign, IL 61820. E-mail: roisman@uiuc.edu.*

Manuscript received 12 Aug. 2007; revised 30 Dec. 2007; accepted 25 Jan. 2008; published online 27 Feb. 2008.
Recommended for acceptance by T. Darrell.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2007-08-0496.
Digital Object Identifier no. 10.1109/TPAMI.2008.52.

0162-8828/09/$25.00 © 2009 IEEE Published by the IEEE Computer Society

5. the multimodal computer-aided learning system[1] in the Beckman Institute, University of Illinois, Urbana-Champaign (UIUC), where the computer avatar offers an appropriate tutoring strategy based on the information of the user's facial expression, keywords, eye movement, and task state.

These systems represent initial efforts toward the future human-centered multimodal HCI.

Except in standard HCI scenarios, potential commercial applications of automatic human affect recognition include affect-sensitive systems for customer services, call centers, intelligent automobile systems, and game and entertainment industries. These systems will change the ways in which we interact with computer systems. For example, an automatic service call center with an affect detector would be able to make an appropriate response or pass control over to human operators [83], and an intelligent automobile system with a fatigue detector could monitor the vigilance of the driver and apply an appropriate action to avoid accidents [69].

Another important application of automated systems for human affect recognition is in affect-related research (e.g., in psychology, psychiatry, behavioral science, and neuroscience), where such systems can improve the quality of the research by improving the reliability of measurements and speeding up the currently tedious manual task of processing data on human affective behavior [47]. The research areas that would reap substantial benefits from such automatic tools include social and emotional development research [111], mother-infant interaction [29], tutoring [54], psychiatric disorders [45], and studies on affective expressions (e.g., deception) [65], [47]. Automated detectors of affective states and moods, including fatigue, depression, and anxiety, could also form an important step toward personal wellness and assistive technologies [100].

Because of this practical importance and the theoretical interest of cognitive scientists, automatic human affect analysis has attracted the interest of many researchers in the last three decades. Suwa et al. [127] presented an early attempt in 1978 to automatically analyze facial expressions. The vocal emotion analysis has an even longer history, starting with the study of Williams and Stevens in 1972 [145]. Since the late 1990s, an increasing number of efforts toward automatic affect recognition were reported in the literature. Early efforts toward machine affect recognition from face images include those of Mase [90], and Kobayashi and Hara [76] in 1991. Early efforts toward the machine analysis of basic emotions from vocal cues include studies like that of Dellaert et al. in 1996 [33]. The study of Chen et al. in 1998 [22] represents an early attempt toward audiovisual affect recognition. For exhaustive surveys of the past work in the machine analysis of affective expressions, readers are referred to [115], [31], [102], [49], [96], [105], [130], [121], and [98], which were published in 1992 to 2007, respectively.

Overall, most of the existing approaches to automatic human affect analysis are the following:

- approaches that are trained and tested on a deliberately displayed series of exaggerated affective expressions,
- approaches that are aimed at recognition of a small number of prototypical (basic) expressions of emotion (i.e., happiness, sadness, anger, fear, surprise, and disgust), and
- single-modal approaches, where information processed by the computer system is limited to either face images or the speech signals.

Accordingly, reviewing the efforts toward the single-modal analysis of artificial affective expressions have been the focus in the previously published survey papers, among which the papers of Cowie et al. in 2001 [31] and of Pantic and Rothkrantz in 2003 [102] have been the most comprehensive and widely cited in this field to date. At the time when these surveys were written, most of the available data sets of affective displays were small and contained only deliberate affective displays (mainly of the six prototypical emotions) recorded under highly constrained conditions. Multimedia data were rare, and there was no 3D data on facial affective behavior, there was no data of combined face and body displays of affective behavior, and it was rare to find data that included spontaneous displays of affective behavior.

Hence, while automatic detection of the six basic emotions in posed controlled audio or visual displays can be done with reasonably high accuracy, detecting these expressions or any expression of human affective behavior in less constrained settings is still a very challenging problem due to the fact that deliberate behavior differs in visual appearance, audio profile, and timing from spontaneously occurring behavior. Due to this criticism received from both cognitive and computer scientists, the focus of the research in the field started to shift to the automatic analysis of spontaneously displayed affective behavior. Several studies have recently emerged on the machine analysis of spontaneous facial expressions (e.g., [10], [28], [135], and [4]) and vocal expressions (e.g., [12] and [83]).

Also, it has been shown by several experimental studies that integrating the information from audio and video leads to an improved performance of affective behavior recognition. The improved reliability of audiovisual approaches in comparison to single-modal approaches can be explained as follows: Current techniques for the detection and tracking of facial expressions are sensitive to head pose, clutter, and variations in lighting conditions, while current techniques for speech processing are sensitive to auditory noise. Audiovisual fusion can make use of the complementary information from these two channels. In addition, many psychological studies have theoretically and empirically demonstrated the importance of the integration of information from multiple modalities (vocal and visual expression in this paper) to yield a coherent representation and inference of emotions [1], [113], [117]. As a result, an increased number of studies on audiovisual human affect recognition have emerged in recent years (e.g., [17], [53], and [151]).

This paper introduces and surveys these recent advances in the research on human affect recognition. In contrast to

1. http://itr.beckman.uiuc.edu.

previously published survey papers in the field, it focuses on the approaches that can handle audio and/or visual recordings of *spontaneous* (as opposed to posed) displays of affective states. It also examines the state-of-the-art methods that have not been reviewed in previous survey papers but are important, specifically for advancing human affect sensing technology. Finally, we discuss the collection and availability of training and test data in detail. This paper is organized as follows: Section 2 describes the human perception of affect from a psychological perspective. Section 3 provides a detailed review of the related studies, including multimedia emotion databases and existing human affect recognition methods. Section 4 discusses some of the challenges that researchers face in this field. A summary and closing remarks conclude this paper.

2 Human Affect (Emotion) Perception

Automatic affect recognition is inherently a multidisciplinary enterprise involving different research fields, including psychology, linguistics, computer vision, speech analysis, and machine learning. There is no doubt that the progress in automatic affect recognition is contingent on the progress of the research in each of those fields [44].

2.1 The Description of Affect

We begin by briefly introducing three primary ways that affect has been conceptualized in psychological research. Research on the basic structure and description of affect is important in that these conceptualizations provide information about the affective displays that automatic emotion recognition systems are designed to detect.

Perhaps the most long-standing way that affect has been described by psychologists is in terms of discrete categories, an approach that is rooted in the language of daily life [40], [41], [46], [131]. The most popular example of this description is the prototypical (basic) emotion categories, which include happiness, sadness, fear, anger, disgust, and surprise. This description of basic emotions was specially supported by the cross-cultural studies conducted by Ekman [40], [42], indicating that humans perceive certain basic emotions with respect to facial expressions in the same way, regardless of culture. This influence of a basic emotion theory has resulted in the fact that most of the existing studies of automatic affect recognition focus on recognizing these basic emotions. The main advantage of a category representation is that people use this categorical scheme to describe observed emotional displays in daily life. The labeling scheme based on category is very intuitive and thus matches people's experience. However, discrete lists of emotions fail to describe the range of emotions that occur in natural communication settings. For example, although prototypical emotions are key points of emotion reference, they cover a rather small part of our daily emotional displays. Selection of affect categories that can describe the wide variety of affective displays that people show in daily interpersonal interactions needs to be done in a pragmatic and context-dependent manner [102], [105].

An alternative to the categorical description of human affect is the dimensional description [58], [114], [140], where an affective state is characterized in terms of a small number of latent dimensions rather than in terms of a small number of discrete emotion categories. These dimensions include evaluation, activation, control, power, etc. In particular, the evaluation and activation dimensions are expected to reflect the main aspects of emotion. The evaluation dimension measures how a human feels, from positive to negative. The activation dimension measures whether humans are more or less likely to take an action under the emotional state, from active to passive. In contrast to categorical representation, dimensional representation enables raters to label a range of emotions. However, the projection of the high-dimensional emotional states onto a rudimentary 2D space results, to some degree, in the loss of information. Some emotions become indistinguishable (e.g., fear and anger), and some emotions lie outside the space (e.g., surprise). This representation is not intuitive, and raters need special training to use the dimensional labeling system (e.g., the Feeltrace system [30]). In automatic emotion recognition systems that are based on the 2D dimensional emotion representation (e.g., [17] and [53]), the problem is often further simplified to two-class (positive versus negative and active versus passive) or four-class (quadrants of 2D space) classification.

One of the most influential emotion theories in modern psychology is the appraisal-based approach [117], which can be regarded as the extension of the dimensional approach described above. In this representation, an emotion is described through a set of stimulus evaluation checks, including the novelty, intrinsic pleasantness, goal-based significance, coping potential, and compatibility with standards. However, translating this scheme into one engineering framework for purposes of automatic emotion recognition remains challenging [116].

2.2 Association between Affect, Audio, and Visual Signals

Affective arousal modulates all human communicative signals. Psychologists and linguists have various opinions about the importance of different cues (audio and visual cues in this paper) in human affect judgment. Ekman [41] found that the relative contributions of facial expression, speech, and body gestures to affect judgment depend both on the affective state and the environment where the affective behavior occurs. On the other hand, some studies (e.g., [1] and [92]) indicated that a facial expression in the visual channel is the most important affective cue and correlates well with the body and voice. Many studies have theoretically and empirically demonstrated the advantage of the integration of multiple modalities (vocal and visual expression) in human affect perception over single modalities [1], [113], [117].

Different from the traditional message judgment, in which the aim is to infer what underlies a displayed behavior such as affect or personality, another major approach to human behavior measurement is the sign judgment [26]. The aim of sign judgment is to describe the appearance, rather than the meaning, of the shown behavior such as facial signal, body gesture, or speech rate. While message judgment is focused on interpretation, sign judgment attempts to be an objective description, leaving the inference about the conveyed message to high-level decision making. As indicated by Cohn [26], the most commonly used sign judgment method for the manual labeling of facial behavior is the Facial Action Coding System (FACS) proposed by Ekman et al. [43]. FACS is a comprehensive and anatomically based system that is used to measure all visually discernible facial movements in terms of atomic facial actions called Action Units (AUs). As AUs are independent of interpretation, they can be used for any high-level decision-making process, including the recognition of basic emotions according to Emotional FACS (EMFACS) rules[2], the recognition of various affective states according to the FACS Affect Interpretation Database (FACSAID)[2] introduced by Ekman et al. [43], and the recognition of other complex psychological states such as depression [47] or pain [144]. AUs of the FACS are very suitable to use in studies on human naturalistic facial behavior, as the thousands of anatomically possible facial expressions (independent of their high-level interpretation) can be described as combinations of 27 basic AUs and a number of AU descriptors. It is not surprising, therefore, that an increasing number of studies on human spontaneous facial behavior are based on automatic AU recognition (e.g., [10], [27], [135], [87], and [134]).

Speech is another important communicative modality in human-human interaction. Speech conveys affective information through explicit (linguistic) and implicit (paralinguistic) messages that reflect the way that the words are spoken. As the linguistic content is concerned, some information about the speaker's affective state can be inferred directly from the surface features of words, which were summarized in some affective word dictionaries and lexical affinity [110], [142], and the rest of affective information lies below the text surface and can only be detected when the semantic context (e.g., discourse information) is taken into account. However, findings in basic research [1], [55] indicate that linguistic messages are rather unreliable means of analyzing human (affective) behavior, and it is very difficult to anticipate a person's word choice and the associated intent in affective expressions. In addition, the association between linguistic content and emotion is language dependent, and generalizing from one language to another is very difficult to achieve.

When it comes to implicit paralinguistic messages that convey affective information, basic researchers have not identified an optimal set of voice cues that reliably discriminate among emotions. Nonetheless, listeners seem to be accurate in decoding some basic emotions from prosody [70] and some nonbasic affective states such as distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, and yawns [113]. Cowie et al. [31] provided a comprehensive summary of qualitative acoustic correlations for prototypical emotions.

In summary, a large number of studies in psychology and linguistics confirm the correlation between some affective displays (especially prototypical emotions) and specific audio and visual signals (e.g., [1], [47], and [113]). The human judgment agreement is typically higher for facial expression modality than for vocal expression modality. However, the amount of the agreement drops considerably when the stimuli are spontaneously displayed expressions of affective behavior rather than posed exaggerated displays. In addition, facial expression and the vocal expression of emotion are often studied separately. This precludes finding evidence of the temporal correlation between them. On the other hand, a growing body of research in cognitive sciences argues that the dynamics of human behavior are crucial for its interpretation (e.g., [47], [113], [116], and [117]). For example, it has been shown that temporal dynamics of facial behavior represent a critical factor for distinction between spontaneous and posed facial behavior (e.g., [28], [47], [135], and [134]) and for categorization of complex behaviors like pain, shame, and amusement (e.g., [47], [144], [4], and [87]). Based on these findings, we may expect that the temporal dynamics of each modality (facial and vocal) and the temporal correlations between the two modalities play an important role in the interpretation of human naturalistic audiovisual affective behavior. However, these are virtually unexplored areas of research.

Another largely unexplored area of research is that of context dependency. The interpretation of human behavioral signals is context dependent. For example, a smile can be a display of politeness, irony, joy, or greeting. To interpret a behavioral signal, it is important to know the context in which this signal has been displayed, i.e., where the expresser is (e.g., inside, on the street, or in the car), what the expresser's current task is, who the receiver is, and who the expresser is [113].

3 THE STATE OF THE ART

Rather than providing exhaustive coverage of all past efforts in the field of automatic recognition of human affect, we focus here on the efforts recently proposed in the literature that have not been reviewed elsewhere, that represent multimodal approaches to the problem of human affect recognition, that address the problem of the automatic analysis of spontaneous affective behavior, or that represent exemplary approaches to treating a specific problem relevant for achieving a better human affect sensing technology. Due to limitations on space and our knowledge, we sincerely apologize to those authors whose work is not included in this paper. For exhaustive surveys of the past efforts in the field, readers are referred to the following articles:

- Overviews of early work on facial expression analysis: [115], [101], and [49].
- Surveys of techniques for automatic facial muscle action recognition and facial expression analysis: [130] and [98].
- Overviews of multimodal affect recognition methods: [31], [102], [105], [121], [68], and [152] (this is a short preliminary version of the survey presented in this current paper).

In this section, we first offer an overview of the existing databases of audio and/or visual recordings of human affective displays, which provide the basis of automatic

2. http://face-and-emotion.com/dataface/general/homepage.jsp.

affect analysis. Next, we examine available computing methods for automatic human affect recognition.

3.1 Databases

Having enough labeled data of human affective expressions is a prerequisite in designing automatic affect recognizer. Authentic affective expressions are difficult to collect because they are relatively rare, short lived, and filled with subtle context-based changes that make it difficult to elicit affective displays without influencing the results. In addition, manual labeling of spontaneous emotional expressions for the ground truth is very time consuming, error prone, and expensive. This state of affairs makes the automatic analysis of spontaneous emotional expression a very difficult task. Due to these difficulties, most of the existing studies on the automatic analysis of human affective displays have been based on the "artificial" material of deliberately expressed emotions, elicited by asking the subjects to perform a series of emotional expressions in front of a camera and/or the microphone.

However, increasing evidence suggests that deliberate behavior differs in visual appearance, audio profile, and timing from spontaneously occurring behavior. For example, Whissell shows that the posed nature of emotions in spoken language may differ in the choice of words and timing from corresponding performances in natural settings [142]. When it comes to facial behavior, there is a large body of research in psychology and neuroscience demonstrating that spontaneous deliberately displayed facial behavior has differences both in utilized facial muscles and their dynamics (e.g., [47]). For instance, many types of spontaneous smiles (e.g., polite) are smaller in amplitude, longer in total duration, and slower in onset and offset times than posed smiles (e.g., [28], [47], and [134]). Similarly, it has been shown that spontaneous brow actions (AU1, AU2, and AU4 in the FACS system) have different morphological and temporal characteristics (intensity, duration, and occurrence order) than posed brow actions [135]. It is not surprising, therefore, that methods of automated human affect analysis that have been trained on deliberate and often exaggerated behaviors usually fail to generalize to the subtlety and complexity of spontaneous affective behavior.

In addition, most of the current human affect recognizers are evaluated using clear constrained input (e.g., high-quality visual and audio recording, nonoccluded, and front- or profile-view face), which is different from the input coming from a natural setting. In addition, most of the emotion expressions that occur in a realistic interpersonal or HCI are nonbasic emotions [32]. Yet, the majority of the existing systems for human affect recognition aim at classifying the input expression as the basic emotion category (e.g., [31], [102], and [105]).

These findings and the general lack of a comprehensive reference set of audio and/or visual recordings of human affective displays motivated several efforts aimed at the development of data sets that could be used for training and test of automatic systems for human affect analysis. Table 1 lists some noteworthy audio, visual, and audiovisual data resources that were reported in the literature. For each database, we provide the following information:

1. affect elicitation method (i.e., whether the elicited affective displays are posed or spontaneous),
2. size (the number of subjects and available data samples),
3. modality (audio and/or visual),
4. affect description (category or dimension),
5. labeling scheme, and
6. public accessibility.

For other surveys of existing databases of human affective behavior, the readers are referred to [32], [59], and [106].

As far as the databases of deliberate affective behavior are concerned, the following databases need to be mentioned. The Cohn-Kanade facial expression database [71] is the most widely used database for facial expression recognition. The BU-3DFE database of Yin and colleagues [148] contains 3D range data of six prototypical facial expressions displayed at four different levels of intensity. The FABO database of Gunes and Piccardi [63] contains videos of facial expressions and body gestures portraying posed displays of basic and nonbasic affective states (six prototypical emotions, uncertainty, anxiety, boredom, and neutral). The MMI facial expression database [106], [98] is, to our knowledge, the most comprehensive data set of facial behavior recordings to date. It contains both posed expressions and spontaneous expressions of facial behavior. The available recordings of deliberate facial behavior are both static images and videos, where a large part of video recordings were recorded in both the frontal and the profile views of the face. The database represents a facial behavior data repository that is available, searchable, and downloadable via the Internet.[3] Although there are many databases of acted emotional speech,[4] a large majority of these data sets contain unlabeled data, which makes them unsuitable for research on automatic vocal affect recognition. The Banse-Scherer vocal affect database [8] and the Danish Emotional Speech database[5] are the two most widely used databases in the research on vocal affect recognition from acted emotional speech. Finally, the Chen-Huang audiovisual database [21] is, to our knowledge, the largest multimedia database containing facial and vocal deliberate displays of basic emotions and four cognitive states:

1. interest,
2. puzzlement,
3. frustration, and
4. boredom.

The existing data sets of spontaneous affective behavior were collected in one of the following scenarios: human-human conversation, HCI, and use of a video kiosk. Human-human conversation scenarios include face-to-face interviews (e.g., [10], [38], [111], and [65]), phone conversations (e.g., [34]), and meetings (e.g., [15] and AMI[6]). HCI scenarios include Wizard of Oz scenarios (e.g., [13] and SAL[7]) and computer-based dialogue systems (e.g., [83] and

3. http://www.mmifacedb.com/.
4. http://emotion-research.net/wiki/Databases.
5. http://cpk.auc.dk/~tb/speech/Emotions/.
6. http://corpus.amiproject.org/.
7. http://emotion-research.net/toolbox/toolboxdatabase.2006-09-26.5667892524.

TABLE 1
Audio and/or Visual Databases of Human Affective Behavior

References	Elicitation method	Size	A/V	Emotion description	Labeling	Accessibility
Cohn-Kanade (CK) '00 [71]	Posed	210 adults, 3 races; Available: 480 videos	V	Category: 6 basic emotions, and AUs	FACS	Y
Sebe et al. (SD) '04 [123]	Natural: Subjects watched emotion-inducing videos	28 adults	V	Category: Neutral, happy, surprise, disgust	Self-report	N
MMI '05[3] [106], [98]	Posed: static images, videos recorded simultaneously in frontal and profile view; Natural: Children interacted with a comedian. Adults watched emotion-inducing videos	Posed: 61 adults Natural: 11 children and 18 adults. Overall: 3 races Available: 1250 videos, 600 static images	V	Category: 6 basic emotions, single AU and multiple AUs activation	FACS, Observers' judgment	Y
UT Dallas '06 [95]	Natural: Subjects watched emotion-inducing videos	229 adults	V	Category: 6 basic emotions, puzzle, laughter, boredom, disbelief	Observers' judgment	Y
BU-3DFE (BU)'06 [148]	Posed: 3D range data by using 3DMD digitizer.	100 adults Mixed races	V	Category: 6 basic emotions. Four levels of intensity	N/A	Y
FABO face and body gesture [63]	Posed: two cameras to record facial expressions and body gestures respectively	23 adults Mixed races Available: 210 videos	V	Category: 6 basic emotions, neutral, uncertainty, anxiety, boredom	N/A	Y
Banse-Scherer '96 [8]	Posed	6 actors & 6 actresses Available: 1344 audio samples	A	Category: hot/cold anger, panic fear, anxiety, despair, sadness, elation, happiness, interest, boredom, shame, pride, disgust, contempt.	Listeners' judgment	Y
Danish Emotional Speech Database '96[5]	Posed	2 actors & 2 actresses; 2 words, 9 sentences, 2 passages; 10 min of audio data.	A	Category: neutral, surprise, happiness, sadness, anger	Listeners' judgment	Y
ISL meeting corpus '02 [15]	Natural: meeting corpus	18 meetings; Available: data of 5 participants per meeting averagely	A	Category: Positive, neutral, negative [3], [90]	Listeners' judgment	Y
CSC corpus [65]	Natural: subject was motivated to tell the truth and deceive the interviewers in different tasks	32 adults, 15.2 h, 3882 speaking turns, 9687 SUs	A	Deceptive, non-deceptive speech	Self-report	N
Automatic call center (ACC)'05 [83]	Natural: Human-computer dialogue at a commercial call system	1187 calls 7200 utterances	A	Category: Negative, non-negative	listeners' judgment	N
Bank and Stock Service 04 [34]	Natural: human-human dialogue at call center	350 dialogues, 10000 speaking turns	A	Category: fear, anger, stress	Listeners' judgment	N
AIBO database '04 [13]	Natural: children and robot interaction	110 dialogues, 29200 words	A	Category: joyful, emphatic, surprised, ironic, helpless, touchy, angry, bored, motherese, reprimanding, rest	Listeners' judgment	N
Chen-Huang (CH) '00 [21]	Posed	100 adults, 9900 visual and AV expressions	AV	Category: 6 basic emotions, and 4 cognitive states (interest, puzzle, bore, frustration)	N/A	N
Adult Attachment Interview (AAI)'04[111]	Natural: subjects were interviewed to describe the childhood experience	60 adults Each interview last 30-60min	AV	Category: 6 basic emotions, embarrassment, contempt, shame, general positive and negative.	FACS	N
RU-FACS (RU) '05 [10]	Natural: subjects were tried to convince the interviewers they were telling the truth	100 adults	AV	Category: 33 AUs	FACS	N
SAL '05[7]	Induced: subjects interacted with artificial listener with different personalities	24 adults 10h	AV	Dimensional labeling/categorical labeling	FEEL-TRACE	Y
Belfast database (BE) '03 [38]	Natural: clips taken from television and realistic interviews with research team	125 subjects. 209 sequences from TV, 30 from interview	AV	Dimensional labeling/categorical labeling	FEEL-TRACE	Y

A: audio, V: video, AV: audiovisual, N/A: not available, Y: yes, and N: not yet.

[86]). In the video kiosk settings (e.g., [95], [98], and [123]), the subjects' affective reactions are recorded while the subjects are watching emotion-inducing videos.

In most of the existing databases, discrete emotion categories are used as the emotion descriptors. The labels of prototypical emotions are often used, especially in the databases of deliberate affective behavior. In databases of spontaneous affective behavior, coarse affective states like positive versus negative (e.g., [15] and [83]), dimensional descriptions in the evaluation-activation space (e.g., SAL), and some application-dependent affective states are usually used as the data labels. Some typical examples of the used application-dependent affect-interpretative labels (e.g., [95], [63], [13], and [111]) are the following:

1. interest,
2. boredom,
3. confusion,
4. frustration,
5. fatigue,

6. empathy,
7. stress,
8. irony,
9. annoyance,
10. amusement,
11. helplessness,
12. panic,
13. shame,
14. reprehension, and
15. rebelliousness.

As explained above, AUs are very suitable to describe the richness of spontaneous facial behavior, as the thousands of anatomically possible facial expressions can be represented as combination of a few dozens of AUs. Hence, the labeling schemes used to code data include FACS AUs (e.g., [10], [71], [106], [98], and [111]), the Feeltrace system for evaluation-activation dimensional description (e.g., [38] and SAL), self report (e.g., [123] and [65]), and human-observer judgment (e.g., [13], [15], [83], [95], and [98]).

The current situation of emotion database research is considerably different from what was described in the comprehensive surveys written by Pantic and Rothkrantz [102] and Cowie et al. [31]. The current state of the art is advanced and can be summarized as follows:

- A database of 3D recordings of acted facial affect [148] and a database of face-and-body recordings of acted affective displays [63] have been made available.
- A collection of acted facial affect displays made from the profile view is shared on the Internet [106].
- Several large audio, visual, and audiovisual sets of human spontaneous affective behavior have been collected, some of which are released for public use.

The existence of these data sets of spontaneous affective behavior is very promising, and we expect that this will produce a major shift in the course of the research in the field: from the analysis of exaggerated expressions of basic emotions to the analysis of naturalistic affective behavior. We also expect subsequent shifts in research in various related fields such as ambient intelligence, transportation, and personal wellness technologies.

3.2 Vision-Based Affect Recognition

Because of the importance of face in emotion expression and perception, most of the vision-based affect recognition studies focus on facial expression analysis. We can distinguish two main streams in the current research on the machine analysis of facial expressions [26], [98]: the recognition of affect and the recognition of facial muscle action (facial AUs). As explained above, facial AUs are a relatively objective description of facial signals and can be mapped to the emotion categories based on a high-level mapping such as EMFACS and FACSAID or to any other set of high-order interpretation categories, including complex affective states like depression [47] or pain [144].

As far as automatic facial affect recognition is concerned, most of the existing efforts studied the expressions of the six basic emotions due to their universal properties, their marked reference representation in our affective lives, and the availability of the relevant training and test material (e.g., [71]). There are a few tentative efforts to detect nonbasic affective states from deliberately displayed facial expressions, including fatigue [60], [69], and mental states like agreeing, concentrated, interested, thinking, confused, and frustrated (e.g., [48], [72], [73], [129], and [147]).

Most of the existing works on the automatic facial expression recognition are based on deliberate and often exaggerated facial displays (e.g., [130]). However, several efforts have been recently reported on the automatic analysis of spontaneous facial expression data (e.g., [9], [10], [11], [27], [28], [67], [88], [123], [135], [149], [87], [4], and [134]). Some of them study the automatic recognition of AUs, rather than emotions, from spontaneous facial displays (e.g., [9], [10], [11], [27], [28], [135], and [134]). Studies reported in [28], [135], [134], and [87] investigated explicitly the difference between spontaneous and deliberate facial behavior. In particular, the studies of Valstar et al. [135], [134] and the study of Littlewort et al. [87] are the first reported efforts to date to automatically discern posed from spontaneous facial behavior. It is interesting to note that, confirming with research findings in psychology (e.g., [47]), the systems proposed by Valstar et al. were built to characterize temporal dynamics of facial actions and employ parameters like speed, intensity, duration, and the co-occurrence of facial muscles activations to classify facial behavior present in a video as either deliberate or spontaneous.

Some of the studies on the machine analysis of spontaneous facial behavior were conducted using the data sets listed in Table 1 (e.g., [10], [149], and [134]). For other studies, new data sets were collected. Overall, the utilized data were collected in the following data-elicitation scenarios: human-human conversation (e.g., [10], [11], [28], [135], [149], and [4]), Wizard of Oz scenarios (e.g., [67]), or TV broadcast (e.g., [147]). The studies reported in [123] and [147] explored the automatic recognition of a subset of basic emotional expressions. The study of Zeng et al. [149] investigated separating emotional state from nonemotional states during the Adult Attachment Interview. Studies on separating posed from genuine smiles were reported in [28] and [134], and studies on the recognition of pain from facial behavior were reported in [4] and [87].

Most of the existing facial expression recognizers employ various pattern recognition approaches and are based on 2D spatiotemporal facial features. The usually extracted facial features are either geometric features such as the shapes of the facial components (eyes, mouth, etc.) and the location of facial salient points (corners of the eyes, mouth, etc.) or appearance features representing the facial texture, including wrinkles, bulges, and furrows. Typical examples of geometric-feature-based methods are those of Chang et al. [19], who used a shape model defined by 58 facial landmarks, of Pantic et al. [98], [99], [103], [135], [134], who used a set of facial characteristic points around the mouth, eyes, eyebrows, nose, and chin, and of Kotsia and Pitas [77], who used the Candide grid. Typical examples of appearance-feature-based methods are listed as follows:

1. Bartlett et al. [9], [10], [11], [87] and Guo and Dyer [64], who used Gabor wavelets,
2. Whitehill and Omlin [143], who used Haar features,

3. Anderson and McOwen [2], who used a holistic spatial ratio face template,
4. Valstar et al. [136], who used temporal templates, and
5. Chang et al. [18], who built a probabilistic recognition algorithm based on the manifold subspace of aligned face appearances.

As suggested in several studies (e.g., [99]), using both geometric and appearance features might be the best choice for designing automatic facial expression recognizers. Typical examples of hybrid geometric and appearance-feature-based methods are those proposed by Tian et al. [130], who used facial component shapes and the transient features like crow-feet wrinkles and nasal-labial furrows, and that of Zhang and Ji [158], who used 26 facial points around the eyes, eyebrows, and mouth, and the transient features proposed by Tian et al. Another example of such a method is that proposed by Lucey et al. [88], who used the Active Appearance Model (AAM) to capture the characteristics of the facial appearance and the shape of facial expressions.

Most of the existing 2D-feature-based methods are suitable for the analysis of facial expressions under a small range of head motions. Thus, most of these methods focus on the recognition of facial expressions in near-frontal-view recordings. An exemplar exception is the study of Pantic and Patras [99], who explored automatic analysis of facial expressions from the profile-view face.

A few approaches to automatic facial expression analysis are based on 3D face models. Huang and colleagues (i.e., [25], [123], [141], and [149]) used features extracted by a 3D face tracker called the Piecewise Bezier Volume Deformation Tracker [128]. Cohn et al. [27] focused on the analysis of brow AUs and head movement based on a cylindrical head model [146]. Chang et al. [20] and Yin et al. [139], [148] used 3D expression data for facial expression recognition. The progress of the methodology based on 3D face models may yield view-independent facial expression recognition, which is important for spontaneous facial expression recognition, because the subject can be recorded in less controlled real-world settings.

Some efforts are reported to decompose multiple factors (e.g., the facial expression, face style, or pose) from face images. Typical examples are those of Wang and Ahuja [137], who used a multilinear subspace method, and of Lee and Elgammal [81], who proposed decomposable nonlinear manifold to estimate facial expression and face style simultaneously. The study of Zhu and Ji [160] used a normalized SVD decomposition to recover facial expression and pose.

Relatively few studies investigated the fusion of the information from facial expressions and head movement (e.g., [27], [69], [158], [160], and [134]), the fusion of facial expression and body gesture (e.g., [7], [61], [62], and [134]), and the fusion of facial expressions and postures from a sensor chair (e.g., [72] and [73]), with the aim of improving affect recognition performance.

Finally, virtually all present approaches to automatic facial expression analysis are context insensitive. Exceptions to this overall state of the art in the field include just a few studies. For example, Pantic and Rothkrantz [104] and Fasel et al. [50] investigated the interpretation of facial expressions in terms of user-defined interpretation labels. Ji et al. [69] investigated the influence of context (work condition, sleeping quality, circadian rhythm, environment, and physical condition) on fatigue detection, and Kapoor and Picard [73] investigated the influence of the task states (difficulty level and game state) on interest detection.

Table 2 provides an overview of the currently existing exemplar systems for vision-based affect recognition with respect to the utilized facial features, classifier, and performance. While summarizing the performance of the surveyed systems, we also mention a number of relevant aspects, including the following:

1. type of the utilized data (spontaneous or posed, the number of different subjects, and sample size),
2. whether the system is person dependent or independent,
3. whether it performs in a real-time condition,
4. what the number of target classification categories is,
5. whether and which other cues, aside from the face, have been used in the classification (head, body, eye, posture, task state, and other contexts),
6. whether the system processes still images or videos, and
7. how accurately it performs the target classification.

A missing entry means that the matter at issue was not reported or it remained unclear from the available literature. For instance, some studies did not explicitly indicate whether the recordings of the same subjects were used as both the testing data and the training data. Hence, it remains unclear whether these systems perform in a subject-independent manner. It is important to stress that we cannot rank the performances of the surveyed systems because each of the relevant studies has been conducted under different experimental conditions using different data, different testing methods (such as person dependent/independent), and different performance measurements (accuracy, equal error rate, etc.).

The research on the machine analysis of facial affect has seen a lot of progress when compared to that described in the survey paper of Pantic and Rothkrantz [102]. The current state of the art in the field is listed as follows:

- Methods have been proposed to detect attitudinal and nonbasic affective states such as confusion, boredom, agreement, fatigue, frustration, and pain from facial expressions (e.g., [69], [72], [129], [147], and [87]).
- Initial efforts were conducted to analyze and automatically discern posed (deliberate) facial displays from genuine (spontaneous) displays (e.g., [135] and [134]).
- First attempts are reported toward the vision-based analysis of spontaneous human behavior based on 3D face models (e.g., [123] and [149]), based on fusing the information from facial expressions and head gestures (e.g., [27] and [134]), and based on

TABLE 2
Vision-Based Affect Recognition

References	Facial Feature	Classifier	Performance							
			exp	per	cues	rea	class	sub	samp	acc (%)
Ashraf et al. 07 [4]	AAM	SVM	S	I		N	2	21	?	Im: EER:19%
Bartlett et al. 04 [9]	Gabor wavelets	SVM+HMM	S	I		N	3 AUs	17	Vi: 230+ (OD)	Im/Vi: 75-98
Bartlett et al. 05 [10], [11],	Gabor wavelets	Adaboost SVM	S, P	I		Y	17 AUs	CK+EH: 119, RU:12	Im: 2568(CK+EH) 1689 (RU)	Im: 93.4(CK+EH), 90.5 (RU)
Cohen et al. 03 [25]	12 motion units	Tree-augmented DBN, HMM	P	D, I		Y	6	CH:5 CK:53	Vi: 30 (CH), 53 (CK)	Im: 66.53(CH), 73.22(CK) Vi: 58.63(CH)
Cohn et al. 04 [27]	shape models, Gabor wavelets	LDC	S	I	H	N	3 AUs	21	Im: 99 (OD)	Im: 76 (3-class)
El Kaliouby & Robinson 04 [48]	24 facial points	DBN	P	D	H	Y	6	30	Vi: 164 (OD)	Vi: 77.4
Fasel et al. 04 [50]	Gray-level intensity	NN	P	?	C	?	7	?	Im: 503 (CK)	Im: 38-68
Gunes & Piccardi 05 [61]	Shape features, optical flow	C4.5, Bayes-Net	P	?	B	N	8	FABO:4	Im: 206 (FABO)	Im: 80-100 (various fusion)
Ioannou et al. 05 [67]	FAPs	neurofuzzy network	S	I		N	3	?	Im: 984 (OD)	Im: 78
Ji et al. 06 [69]	Shape features	DBN	S	?	H,E,C	Y	2	8	Vi: 320min (OD)	Correlation coefficient: 95.3
Kapoor & Picard 05 [73]	Facial and head gesture	GP, SVM HMM, NN	S	?	E, P, T	?	2	8	Vi: 136 (OD)	Vi: 86
Kapoor et al. 07 [72]	Pixel difference of mouth region	Same as in [73]	S	I	E P S T M	?	2	24	Vi: 24 (OD)	Vi: 79.17 Baseline: 58
Lee & Elgammal [81]	Pixel intensity of face region	decomposable model	P	I		N	6	CK: 8 OD: 16	Vi: 48 (CK), 80 (OD)	Vi: 39.58 Im: 61.85
Littlewort et al. 07 [87]	Gabor wavelets	Adaboost SVM	S	I		Y	2	26	Vi: 312	Vi: 72
Lucey et al. 07 [88]	AAM	SVM	S,P	I		N	AUs: CK: 15 OD: 4	CK: 100 OD: ?	?	Im: 95 (CK) with 16.66% FAR, 70.47 (OD)
Pantic & Patras 06 [99]	Facial profile points	Rule-based	P	I		N	27 AUs	MMI: 19	Vi: 119 (MMI)	Vi: 86.3
Pantic & Rothkrantz 04 [103]	frontal and profile facial points	Rule-based	P	I		N	32 AUs	MMI: 25	Im: 454 (MMI)	Im: 86
Pantic & Rothkrantz [104]	same as in [103]	Rule-based, case-based	P	I	U	N	9	MMI: 8	Im: 196 (MMI)	Im: 83
Sebe et al. 04 [123]	12 motion units	kNN	S	I			4	CK: 53 SD: 28	Vi: ? (SD), 212+ (CK)	Im: 93 (CK) 95 (SD)
Tong et al. 07 [132]	Gabor wavelets	Adaboost, DBN	P	I		?	14 AUs	CK: 100 OD: 10	Im: 14000 (CK+OD)	Vi: 93.2 (OD), 93.3(CK)
Valstar et al. 04 [136]	Motion history images	SNoW kNN	P	I		N	15 AUs	MMI: 19 CK: 100	Vi: 344 (CK), 253 (MMI)	Vi: 61 (MMI) 68 (CK)
Valstar et al. 06 [135]	8 facial points	gentle boost, SVM	S, P	I		N	2	MMI:27 CK: 32 CD: 65	Vi: 60(MMI) 60(CK),70(OD)	Vi: 90.7
Valstar et al. 07 [134]	20 facial points	GentleSVM-sigmoid	S, P	I	H, B	N	2	MMI: 52	Vi: 100 (P), 102 (S)	Vi: 94%
Wang & Ahuja 03 [137]	Shape and gray-level texture	NN with HOSVD	S	?		?	7	14	Im: 110 (OD)	Im: 84.58
Wang et al. 06 [139]	3D surface labels	LDA	P	I		N	6	BU: 60	Im: 720 (BU)	Im: 83.6
Wen & Huang 03 [141]	Geometric, ratio-image	Exemplars with GMM	P	I		N	4	CK: 47	Im: 2981 (CK)	Im:75.37
Whitehill & Omlin 06 [143]	Haar features	Adaboost	P	I		Y	11 AUs	?	Im: 580 (OD)	Im: 92.35
Yeasin et al. 06 [147]	Pixel intensity of face	kNN + HMM	P, S			N	6	CK: 97 OD:21	Vi: 488 (CK) 108 (OD)	Vi: 90.7 (CK) 72-82 (OD)
Zeng et al. 06 [149]	Texture with LPP	SVDD	S	D		N	2	AAI: 2	Female:7857 Male: 5230	Im: 79(male), 87(female)

*exp: **S**pontaneous/**P**osed expression, per: person **D**ependent/**I**ndependent, Im/Vi: **Im**age/**Vi**deo based, cues: other cues aside from the face (**H**ead/**B**ody/**E**ye/**S**kin/**P**osture/**T**ask state/pressure **M**ouse/**U**ser-defined classes/other**C**ontext), rea: real time (Y: yes, and N: no), class: number of classes, sub: number of subjects, samp: sample size, acc: Accuracy, AUs: AUs corresponding to AU detection, min: minutes, EER: equal error rate, FAR: false acceptance rate, and GP: Gaussian process.*
AAI, BU, CH, CK, FABO, MMI, RU, and SD are the database names listed in Table 1.
EH: the Ekman-Hager database, OD: Other database, and ?: missing entry.

fusing the information from facial expressions and body gestures (e.g., [61]).

- Few attempts have also been made toward the context-dependent interpretation of the observed facial behavior (e.g., [50], [69], [72], and [104]).
- Advanced techniques in feature extraction and classification have been applied and extended in this field. A few real-time robust systems have been built (e.g., [11]) thanks to the advance of relevant techniques such as real-time face detection and object tracking.

3.3 Audio-Based Affect Recognition

Research on vocal affect recognition is also largely influenced by a basic emotion theory. In turn, most of the existing efforts in this direction aim at the recognition of a subset of basic emotions from speech signals. However, a few tentative studies were published recently on the interpretation of speech signals in terms of certain application-dependent affective states. These studies are those of the following:

1. Hirschberg et al. [65] and Graciarena et al. [57], who attempted deception detection,
2. Liscombe et al. [84], who focused on detecting certainty,
3. Kwon et al. [79], who reported on stress detection,
4. Zhang et al. [157], who investigated speech-based analysis of confidence, confusion, and frustration,
5. Batliner et al. [12], who aimed at detecting trouble,
6. Ang et al. [3], who explored speech-based recognition of annoyance and frustration, and
7. Steidl et al. [125], who conducted studies on the detection of empathy.

In addition, few efforts toward the automatic recognition of nonlinguistic vocalizations like laughters [133], coughs [91] and cries [97] have also been reported recently. This is of particular importance for the research on the machine analysis of human affects since recent studies in cognitive sciences showed that listeners seem to be rather accurate in decoding some nonbasic affective states such as distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, and yawns [113].

Most of the existing systems for automatic vocal affect recognition were trained and tested on speech data that was collected by asking actors to speak prescribed utterances with certain emotions (e.g., [6] and [79]). As the utterances are isolated from the interaction context, this experimental strategy precludes finding and using correlations between the paralinguistic displays and the linguistic content, which seem to play an important role for affect recognition in daily interpersonal interactions.

Based on the above consideration, researchers started to focus on affect recognition in naturalistic audio recordings collected in call centers (e.g., [35], [82], [83], and [94]), meetings (e.g., [94]), Wizard of Oz scenarios (e.g., [12]), interviews (e.g., [65]), and other dialogue systems (e.g., [14] and [86]). In these natural interaction data, affect displays are often subtle, and basic emotion expressions seldom occur. It is therefore not surprising that recent studies in the field, which are based on such data, attempt to detect either coarse affective states, i.e., positive, negative, and neutral states (e.g., [82], [83], [86], and [94]) or application-dependent states mentioned above rather than basic emotions.

Most of the existing approaches to vocal affect recognition used acoustic features as classification input based on the acoustic correlation for emotion expressions that were summarized in [31]. The popular features are prosodic features (e.g., pitch-related feature, energy-related features, and speech rate) and spectral features (e.g., MFCC and cepstral features). Many studies show that pitch and energy among these features contribute the most to affect recognition (e.g., [79]). An exemplar effort is that of Vasilescu and Devillers [36], who show the relevance of speech disfluencies (e.g., filler and silence pauses) to affect recognition.

With the research shift toward the analysis of spontaneous human behavior, the analysis of acoustic information will not only suffice for identifying subtle changes in vocal affect expression. As indicated by Batliner et al. [12], "The closer we get to a realistic scenario, the less reliable is prosody as an indicator of the speaker's emotional state." In the preliminary experiments of Devillers and Vidrascu [35], using lexical cues resulted in a better performance than using paralinguistic cues to detect relief, anger, fear, and sadness in human-human medical call conversations. In turn, several studies investigated the combination of acoustic features and linguistic features (language and discourse) to improve vocal affect recognition performance. Typical examples of linguistic-paralinguistic fusion methods are those of the following:

1. Litman and Forbes-Riley [86] and Schuller et al. [120], who used spoken words and acoustic features,
2. Lee and Narayanan [83], who used prosodic features, spoken words and information of repetition,
3. Graciarena et al. [57], who combined prosodic, lexical and cepstral features, and
4. Bartliner et al. [12], who used prosodic features, part of speech (POS), dialogue act (DA), repetitions, corrections, and syntactic-prosodic boundary to infer the emotion.

Litman et al. [86] and Forbes-Riley and Litman [52] also investigated the role of the context information (e.g., subject, gender, and turn-level features representing local and global aspects of the dialogue) on audio affective recognition.

Although the above studies indicated recognition improvement by using information on language, discourse, and context, the automatic extraction of these related features is a difficult problem. First, existing automatic speech recognition (ASR) systems cannot reliably recognize the verbal content of emotional speech (e.g., [5]). Second, extracting semantic discourse information is even more challenging. Most of these features are typically extracted manually or directly from transcripts.

Table 3 provides an overview of the currently existing exemplar systems for audio-based affect recognition with respect to the utilized auditory features, classifier, and performance. As in Table 2, we specify relevant aspects in Table 3 to summarize the reported performance of surveyed systems.

The current state of the art in the research field of automatic audio-based affect recognition can be summarized as follows:

- Methods have been proposed to detect nonbasic affective states, including coarse affective states such as negative and nonnegative states (e.g., [83]), application-dependent affective states (e.g., [3], [12], [65], [79], [157], and [125]), and nonlinguistic vocalizations like laughter and cry (e.g., [133], [91], and [97]).
- A few efforts have been made to integrate paralinguistic features and linguistic features such as lexical, dialogic, and discourse features (e.g., [12], [35], [57], [83], [86], and [120]).
- Few investigations have been conducted to make use of contextual information to improve the affect recognition performance (e.g., [52] and [86]).
- Few reported studies have analyzed the affective states across languages (e.g., [94] and [133]).

TABLE 3
Audio-Based Affect Recognition

References	Feature	Classifier	Performance							
			exp	per	cont	class	sub	samp	acc (%)	other
Ang et al. 02 [3]	Prosody, LM features, position, repeats/ correction	Decision tree	S	I		2	837	21899	64-93	Various label and feature conditions
Austermann et al. 05 [6]	Prosody	Fuzzy rules	S	D, I		5	D: 4 I: 4	D: 280 I: 260	D: 84 I: 60	Robot head data
Batliner et al. 03 [12]	prosody, POS, DA, repetitions, corrections, etc.	MLP, LDA	S, P	I		2	A: 1 R: 19 W: 24	A: 10316 R: 13053 W: 28649	A: 95.7 R: 79.6 W: 74.2	AIBO data
Devillers & Vasilescu 06 [35]	Lexical cues, prosody, spectrum, disfluency, etc.	SVM	S	I		4	680	2258	Lexical: 78 paralinguistic: 60	Medical emergency center data
Forbes-Riley & Litman 04 [52]	prosodic, lexical, syntactic, dialogue features, etc.	boost decision tree	S	I	Su, G, T	3	17	453	84.75	computer tutor data
Graciarena et al. 06 [57]	Prosodic, acoustic, lexical	SVM, GMM	S	D		2	32	9328	64.4	CSC data
Hirschberg et al. 05 [65]	Prosodic, acoustic, lexical	Ripper rule-induction	S	D	Dep	2	32	9491	66.4	CSC data
Kwon et al. 03 [79]	Prosody, MFCC	QDA, SVM, HMM LDA	P	D, I		2, 4, 5	OD: 9; AIBO: 14	OD: 8820; AIBO: 3534	2 class: 96 4 class: 70.1 5 class: 42.3	AIBO and OD data
Lee & Narayanan 03 [82]	Prosody	Fuzzy inference	S	I		2	?	F: 776; M: 591	F: 73 M: 63	
Lee & Narayanan 05 [83]	prosody, lexical, and discourse	LDC, kNN	S	I	G	2	ACC: 1187 calls	7200	M: 89.55 F: 92.1	M: 76.5%BL F: 74.1%BL
Liscombe et al. 05 [84]	Acoustic-prosodic	C4.5 with Adaboost	S	I		3	17	6778 turns	76.42	60% BL
Litman & Forbes-Riley 04 [86]	Acoustic-prosodic, lexical	Boost decision tree	S	I	Su, G, T	2, 3	10	333	NPN: 47-67 NnN: 64-72 EnE: 52-75	Various label and feature conditions
Matos et al. 06 [91]	MFCC	HMM	S	I		2	19	Train: 2473 Test:2155	82	
Neiberg et al. 06 [94]	MFCC, MFCC-low, pitch	GMM	S	I		3		OD: 7619 ISL: 12479	OD: 90 ISL: 80	Swedish, English
Schuller et al. 05 [120]	Acoustic-Prosodic, linguistic	StackingC MLR, NB, ND SVM, C4.5	S, P	D, I		7	13+	4336	I: 76.4 D: 94.8	
Steidl et al. 05 [125]	Prosodic, POS features	?	S	I		4	AIBO: 51	6071	60	Entropy measure
Truong & van Leeuwen 07 [133]	Spectral, prosodic	GMM+ SVM	S	I		2	OD1: 34 OD2: 8	OD1: 6838 OD2: 335	EER: 2.9-7.5	English, Dutch
Vasilescu & Devillers 05 [36]	prosodic, spectral, disfluency, etc.	SVM, logistic model tree	S	I	R	2	404	800	82	Same database as [35]
Zhang et al. 04 [157]	Lexical, prosodic, spectral, syntactic	CART tree	S	I		3	OD: 17	714	91.3	

*exp: **S**pontaneous/**P**osed expression, per: person **D**ependent/**I**ndependent, cont: contextual information (**Su**bject/**G**ender/**T**ask/Speaker**R**ole/Speaker-**Dep**endentFeature), class: the number of classes, sub: the number of subjects. samp: sample size (the number of utterances), acc: accuracy, ?: missing entry, BL: Baseline, EER: equal error rate, NPN: negative/neutral/positive, NnN: negative/nonnegative, EnE: emotional/nonemotional, M: male, F: female, A: actor data, R: reading data, W: data of Wizard of Oz, and OD: other database.*
ACC, AIBO, CSC, and ISL are the database names listed in Table 1.

- Some studies have investigated the influence of ambiguity of human labeling on recognition performance (e.g., [3] and [86]) and proposed measures of comparing human labelers and machine classifiers (e.g., [125]).
- Advanced techniques in feature extraction, classification, and natural language processing have been applied and extended in this field. Some studies have been tested on commercial call data (e.g., [83] and [35]).

3.4 Audiovisual Affect Recognition

In the survey of Pantic and Rothkrantz [102], only four studies were found to focus on audiovisual affect recognition. Since then, an increasing number of efforts are reported in this direction. Similar to the state of the art in single-modal affect recognition, most of the existing audiovisual affect recognition studies investigated the recognition of the basic emotions from deliberate displays. Relatively few efforts have been reported toward the detection of nonbasic affective states from deliberate displays. Those include the work of Zeng et al. [150], [153], [154], [155] and that of Sebe et al. [122], who added four cognitive states, considering the importance of these cognitive states in HCI:

1. interest,
2. puzzlement,
3. frustration, and
4. boredom.

Related studies conducted on naturalistic data include that of Pal et al. [97], who designed a system for detecting hunger and pain, as well as sadness, anger, and fear, from infant facial expressions and cries, and that of Petridis and Pantic [108], who investigated separating speech from laughter episodes based on both facial and vocal expression.

Most of the existing methods for audiovisual affect analysis are based on deliberately posed affect displays (e.g., [16], [56], [66], [122], [124], [138], [150], [153], [154], and [155]). Recently, a few exceptional studies have been reported toward audiovisual affect analysis in spontaneous affect displays (e.g., [17], [53], [74], [97], [151], and [108]). Zeng et al. [151] used the data collected in psychological research interview (Adult Attachment Interview), Pal et al. [97] used recordings of infants [97], and Petridis and Pantic [108] used the recordings of people engaged in meetings AMI corpus. On the other hand, Fragopanagos and Taylor [53], Caridakis et al. [17], and Karpouzis et al. [74], used the data collected in Wizard of Oz scenarios. Since the available data were usually insufficient to build a robust machine learning system for the recognition of fine-grained affective states (e.g., basic emotions), the recognition of coarse affective states was attempted in most of the aforementioned studies. The studies of Zeng et al. focus on audiovisual recognition of positive and negative affect [151], while other studies report on the classification of audiovisual input data into the quadrants in the evaluation-activation space [17], [53], [74]. The studies reported in [17], [53], and [74] applied the Feeltrace system that enables raters to continuously label changes in affective expressions. However, note that the study discussed in [53] reported on a considerable labeling variation among four human raters due to the subjectivity of audiovisual affect judgment. More specifically, one of the raters mainly relied on audio information when making judgments, while another rater mainly relied on visual information. This experiment actually also reflects the asynchronization of audio and visual expression. In order to reduce this variation of human labels, the studies of Zeng et al. [151] made the assumption that facial expression and vocal expression have the same coarse emotional states (positive and negative) and then directly used FACS-based labels of facial expressions as audiovisual expression labels.

The data fusion strategies utilized in the current studies on audiovisual affect recognition are feature-level, decision-level, or model-level fusion. Typical examples of feature-level fusion are those reported in [16], [118], and [156], which concatenated the prosodic features and facial features to construct joint feature vectors, which are then used to build an affect recognizer. However, the different time scales and metric levels of features coming from different modalities, as well as increasing feature-vector dimensions influence the performance of an affect recognizer based on a feature-level fusion. The vast majority of studies on bimodal affect recognition reported on decision-level data fusion (e.g., [16], [56], [66], [97], [151], [153], [155], [138], and [108]). In the decision-level data fusion, the input coming from each modality is modeled independently, and these single-modal recognition results are combined in the end. Since humans display audio and visual expressions in a complementary redundant manner, the assumption of conditional independence between audio and visual data streams in decision-level fusion is incorrect and results in the loss of information of mutual correlation between the two modalities. To address this problem, a number of model-level fusion methods have been proposed which aim at making use of the correlation between audio and visual data streams and relaxing the requirement of synchronization of these streams (e.g., [17], [53], [122], [124], [150], and [154]). Zeng et al. [154] presented a Multistream Fused HMM to build an optimal connection among multiple streams from audio and visual channels according to the maximum entropy and the maximum mutual information criterion. Zeng et al. [150] extended this fusion framework by introducing a middle-level training strategy, under which a variety of learning schemes can be used to combine multiple component HMMs. Song et al. [124] presented a tripled HMM to model the correlation properties of three component HMMs that are based individually on upper face, lower face, and prosodic dynamic behaviors. Fragopanagos and Taylor [53] proposed an artificial neural network (NN) with a feedback loop called ANNA to integrate the information from face, prosody, and lexical content. Caridakis et al. [17], Karpouzis et al. [74], and Petridis and Pantic [108] investigated combining the visual and audio data streams by using NNs. Sebe et al. [122] used a Bayesian network (BN) to fuse the facial expression and prosody expression.

Table 4 provides an overview of the currently existing exemplar systems for audiovisual affect recognition with respect to the utilized auditory and visual features, classifier, and performance. As in Tables 2 and 3, we also specify a number of relevant issues in Table 4 to summarize the reported performance of surveyed systems.

In summary, research on audiovisual affect recognition has witnessed significant progress in the last few years as follows:

- Efforts have been reported to detect and interpret nonbasic genuine (spontaneous) affective displays in terms of coarse affective states such as positive and negative affective states (e.g., [151]), quadrants in the evaluation-activation space (e.g., [17], [53], and [74]), and application-dependent states (e.g., [122], [154], [97], and [108]).
- Few studies have been reported on efforts to integrate other affective cues aside from the face and the prosody such as body and lexical features (e.g., [53] and [74]).
- Few attempts have been made to recognize affective displays in specific naturalistic settings (e.g., in a car [66]) and in multiple languages (e.g., [138]).
- Various multimodal data fusion methods have been investigated. In particular, some advanced data fusion methods have been proposed, such as HMM-based fusion (e.g., [124], [154], and [150]), NN-based fusion (e.g., [53] and [74]), and BN-based fusion (e.g., [122]).

4 Challenges

The studies reviewed in Section 3 indicate two new trends in the research on automatic human affect recognition: the

TABLE 4
Audiovisual Affect Recognition

References	Feature	Fusion	Classifier	Performance							
				exp	per	cue	class	sub	samp	acc (%)	other
Busso et al. 04 [16]	102 markers, prosody	F, D	SVM	P	D		4	1	256 sentences	89	
Caridakis et al. 06 [17]	facial points, prosody	M	RNN	S	I		4	SAL 4	1000 tunes	79	
Fragopanagos and Taylor 05 [53]	17 FAPs, prosody	M	ANNA	S	I	L	4	SAL 4	500 epochs	44-71	various labels/ features
Go et al. 03 [56]	Eigenfaces, MFCC	D	LDA	P	I		6	20	360 utterances	95-98	
Hoch et al. 05 [66]	Gabor feature, prosody	D	SVM	P	D		3	7	840 sequences	90.7	car setting
Karpouzis et al. 07 [74]	19 FPs, prosody	M	RNN	S	I	B	4	SAL 4	1000 tunes	82	
Pal et al. 06 [97],	Vertical gray level, F0-F3	D	Rules, k-means	S	D		5	1	?	75.2	
Petridis & Pantic 08 [108]	facial points, cepstral features	F, D	Adaboost + NN	S	I	B	2	8	96 laughter/ speech episodes	86.9	AMI data[6]
Schuller et al. 07 [118]	AAM, prosody, articulatory, voice quality, lexical	F	SVM	S	I	B	3	21	10.5 hours	recall: 41.7-63.9 (RR)	balance training
Sebe et al. 06 [122]	12 motion units, prosody	M	BN	P	D		11	SD 38	1254 sentences	90	
Song et al. 04 [124]	54 FAPs, prosody	M	THMM	P	?		7	?	?	84.7	
Wang & Guan 05 [138]	Gabor wavelets, prosody, MFCC, formants.	D	FLDA	P	I		6	8	500 sentences	82.14	6 lan-guages
Zeng et al. 06 [150]	12 motion units, prosody	M	MFHMM	P	I		11	CH 20	660 sentences	83	
Zeng et al. 07 [151]	Texture with LLP, prosody	D	Adaboost + MHMM	S	D		2	AAI 2	137 uttarances	89	
Zeng et al. 04 [153]	motion units, prosody	D	SNoW	P	D		11	CH 38	1254 sentences	89-90	
Zeng et al. 05 [154]	motion units, prosody	M	MFHMM	P	I		11	CH 20	660 sentences	80.61	
Zeng et al. 07 [155]	motion units, prosody, formants	D	HMM	P	D, I		11	CH 20	660 sentences	I: 72.42 D: 96.3	
Zeng et al. 05 [156]	motion units, prosody	F	Fisher-Boosting	P	D		4	CH 20	660 sentences	84-87	

*Fusion: **F**eature/**D**ecision/**M**odel-level, exp: **S**pontaneous/**P**osed expression, per: person-**D**ependent/**I**ndependent, class: the number of classes, sub: the number of subjects, samp: sample size (the number of utterances), cue: other cues (**L**exical/**B**ody), acc: accuracy, RR: mean with weighted recall values, FAP: facial animation parameter, and ?: missing entry.*
AAI, CH, SAL, and SD are the database names listed in Table 1.

analysis of spontaneous affective behavior and the multimodal analysis of human affective behavior, including audiovisual analysis, combined linguistic and nonlinguistic analysis, and multicue visual analysis based on facial expressions, head movements, and/or body gestures. Several previously recognized problems have been addressed, including the development of more comprehensive data sets of training and testing materials. At the same time, several new challenging issues have been recognized, including the necessity of studying the temporal correlations between the different modalities (audio and visual) and between various behavioral cues (e.g., facial, head, and body gestures). This section discusses these issues in detail.

4.1 Databases

Acquiring valuable spontaneous affective behavior data and the related ground truth is far from being solved. While it is relatively easy to elicit joyful laughter by showing the subjects clips from comedies, the majority of affective states are much more difficult (if possible at all) to elicit (e.g., fear, stress, sadness, or anger, which is particularly difficult to elicit in any laboratory setting, including face-to-face conversation [23]). Social psychology has provided a host of creative strategies for inducing emotion, which seem to be useful for collecting affective expressions that are difficult to elicit in the laboratory and affective expressions that are contextually complex (such as embarrassment) or for research programs that emphasize the "mundane realism" of experimentally elicited emotions [23]. However, engineers, who are usually the designers of the databases of human behavior data, are often not even aware of these strategies, let alone putting them into practice. This situation needs to be changed if the challenging and crucial issue of collecting valuable data on human spontaneous affective behavior is to be addressed.

Although many efforts have been done toward the collection of databases of spontaneous human affective behavior, most of the data contained in the available databases currently lack labels. In other words, no metadata is available which could identify the affective state displayed in a video sample and the context in which this affective state was displayed. There are some related issues.

First, it is not clear which kind of metadata needs to be provided. While data labeling is easy to accomplish in the case of prototypical expressions of emotions, it becomes a real challenge once we move beyond the six basic emotions. To reduce the subjectivity of data labeling, it is generally accepted that human facial expression data need to be FACS coded. The main reason is that FACS AUs are objective descriptors and independent of interpretation and can be used for any high-level decision-making process, including the recognition of affective states. However, while this solves the problem of attaining objective facial behavior coding, how one can objectively code vocal behavior remains an open issue. Nonlinguistic vocalizations like laughter, coughs, cries, etc., can be labeled as such, but there is no set of interpretation-independent codes for labeling emotional speech. Another related issue is that of culture and context dependency. The metadata about the context in which the recordings were made, such as the utilized stimuli, the environment, and the presence of other people, is needed, since these contextual variables may influence masking of the emotional reactions.

Second, even if labeled data are available, engineers responsible for designing an automated human affect analyzer usually assume that the data are accurately labeled. This assumption may or may not be accurate [26], [125]. The reliability of the coding can be ensured by asking several independent human observers to conduct the coding. If the interobserver reliability is high, the reliability of the coding is assured. Interobserver reliability can be improved by providing thorough training to observers on the utilized coding schemes such as FACS. When it comes to data coding in terms of affect labels, a possible method is to use a multilabel multi-time-scale system in order to reduce the subjectivity of human judgment and to represent comprehensive properties of affect displays [37], [80].

Third, human labeling of affective behavior is very time consuming and expensive. In the case of facial expression data, it takes more than 1 hour to manually score 100 still images or 1 minute of video sequence in terms of AUs [43]. A remedy could be the semisupervised active learning method [159], which combines semisupervised learning [24] and active learning [51]. The semisupervised learning mechanism aims at making use of the unlabeled data, and the active learning mechanism aims at enlarging the useful information conveyed by human feedback (annotation in this application) and provides the annotators with the most ambiguous samples according to the current emotion classifier. More specifically, several promising prototype systems were reported in the last few years which can recognize deliberately produced AUs in either (near-) frontal view face images (e.g., [98] and [130]) or profile-view face images (e.g., [99]). Although these systems will not be always able to be generalized to the subtlety and complexity of human affective behavior occurring in real-world settings, they can be used to attain an initial data labeling that can be subsequently controlled and corrected by human observers. However, as this has not been attempted in practice, there is no guarantee that such an approach will actually reduce the time needed for obtaining the ground truth. Future research is needed to determine whether this attempt is feasible.

Although much effort has been done toward the collection of databases of spontaneous human affective behavior, many of these data sets are not publicly available (see Table 1). Some are still under construction, some are in the process of data publication, and some seem to have dim prospects of being published due to the lack of appropriate agreement of subjects. More specifically, spontaneous displays of emotions, especially in audiovisual format, reveal personal and intimate experience; privacy issues jeopardize the public accessibility of many databases.

Aside from these problems concerned with acquiring valuable data, the related ground truth, and the agreement of subjects to make the data publicly available, another important issue is how one can construct and administer such a large affective expression benchmark database. A noteworthy example is the MMI facial expression database [98], [106], which was built as a Web-based direct-manipulation application, allowing easy access and easy search of the available images. In general, in the case of publicly available databases, once the permission for usage is issued, large unstructured files of material are sent. Such unstructured data is difficult to explore and manage. Pantic et al. [102], [106] and Cowie et al. [32] emphasized a number of specific research and development efforts needed to build a comprehensive readily accessible reference set of affective displays that could provide a basis for benchmarks for all different efforts in the research on the machine analysis of human affective behavior. Nonetheless, note that their list of suggestions and recommendations is not exhaustive of worthwhile contributions.

4.2 Vision-Based Affect Recognition

Although several of the efforts discussed in Section 3.2 were recently reported on the machine analysis of spontaneous facial expressions, the problem of the automatic analysis of facial behavior in unconstrained environments is still far from being solved.

Existing methods for the machine analysis of facial affect typically assume that the input data are near frontal- or profile-view face image sequences, showing nonoccluded facial displays captured under constant lighting condition against a static background. In real interaction environments, such an assumption is often invalid. The development of robust face detectors, head, and facial feature trackers, which will be robust to arbitrary head movement, occlusions, and scene complexity like the presence of other people and dynamic background, forms the first step in the realization of facial affect analyzers capable of handling unconstrained environments. View-independent facial expression recognition based on 3D face models (e.g., [20] and [148]) or multiview face models (e.g., [160]) may be a (part of the) solution.

As mentioned already in Section 2, a growing body of research in cognitive sciences argues that the dynamics of human behavior are crucial for its interpretation (e.g., [47] and [113]). For instance, it has been shown that spontaneous smiles are longer in the total duration, can have multiple apexes (multiple rises of the mouth corners), appear before or simultaneously with other facial actions such as the rise of the cheeks, and are slower in the onset and offset times than the posed smiles (e.g., a polite smile) [28]. In spite of these findings, the vast majority of the past work in the field does not take the dynamics of facial expressions into account when analyzing shown facial behavior. Some of the past work in the field has used aspects of temporal structure of facial expression such as the speed of a facial point displacement or the persistence of facial parameters over time (e.g., [87], [132], and [158]). However, just a few recent studies analyze explicitly the temporal structure of facial expressions (e.g., [98], [99], [135], [132], and [134]). In addition, it remains unresolved how the grammar of facial behavior can be learned and how this information can be properly represented and used to handle ambiguities in the input data [100], [102].

Except for a few studies (e.g., [27] and [61]), existing efforts toward the machine analysis of human facial behavior focus only on the analysis of facial gestures without taking into consideration other visual cues like head movements, gaze direction, and body gestures. However, research on cognitive science reports that human judgments of behavioral cues are the most accurate when both the face and the body are taken into account (e.g., [1] and [117]). This seems to be of particular importance when judging certain complex mental states such as embarrassment [75]. However, integration, temporal structures, and temporal correlations between different visual cues are virtually unexplored areas of research. One noteworthy study that investigated fully the automatic coding of human behavior dynamics with respect to both the temporal segments (onset, apex, offset, and neutral) of various visual cues and the temporal correlation between different visual cues (facial, head, and shoulder movements) is that of Valstar et al. [134], who investigated separating posed from genuine smiles in video sequences.

4.3 Audio-Based Affect Recognition

One challenge in audio expression analysis is how we can identify affect-related features in speech signals. When our aim is to detect spontaneous emotion expressions, we have to take into account both linguistic and paralinguistic cues that mingle together in audio channel. Although a number of linguistic and paralinguistic features (e.g., prosodic, dysfluency, lexicon, and discourse features) were proposed in the body of literature on vocal affect recognition, the optimal feature set has not yet been established.

Another challenge is how we can reliably extract these linguistic and paralinguistic features from the audio signals in an automatic way. When prosody is analyzed in a naturalistic conversation, we have to consider the multiple functions of prosody that include information about the expressed affect and a variety of linguistic functions [93]. A prosodic event model that could reflect both linguistic and paralinguistic (affective) functions simultaneously would be an ideal solution. Automatic extraction of spoken words from spontaneous emotional speech is still a difficult problem: the recognition rate of the exiting ASR systems drops significantly as soon as emotional speech needs to be processed. Some tentative studies on adapting an ASR system to emotional speech were reported in [5] and [119]. We hope that, in the future, more such studies will be conducted. In addition, automatic extraction of high-level semantic linguistic information (e.g., DA, repetitions, corrections, and syntactic information) is an even more challenging problem in the research field of natural language processing.

It is interesting to note that some mental states such as frustration and boredom seem to be identifiable from nonlinguistic vocalizations like sighs and yawns [113]. Few efforts toward the automatic recognition of nonlinguistic vocalizations like laughers [133], [108], cries [97], and coughs [91] have also been recently reported. However, no effort toward human affect analysis based on vocal outbursts has been reported so far.

4.4 Audiovisual Affect Recognition

The research on audiovisual affect analysis in naturalistic data is still in its pioneering phase. While all agree that multisensory fusion, including audiovisual data fusion, linguistic and paralinguistic data fusion, and multivisual cue data fusion, would be highly beneficial for the machine analysis of human affect, it remains unclear how this should be accomplished. Studies in neurology on the fusion of sensory neurons [126] are supportive of early data fusion (i.e., feature-level data fusion) rather than of late data fusion (i.e., decision-level fusion). However, it is an open issue how one can construct suitable joint feature vectors composed of features from different modalities with different time scales, different metric levels, and different temporal structures. Simply concatenating audio and video features into a single feature vector, as done in the current human affect analyzers that use feature level data fusion, is obviously not the solution to the problem.

Due to these difficulties, most researchers choose decision-level fusion, in which the input coming from each modality is modeled independently, and these single-modal recognition results are combined in the end. Decision-level fusion, also called classifier fusion, is now an active area in the machine learning and pattern recognition fields. Many studies have demonstrated the advantage of classifier fusion over the individual classifiers due to the uncorrelated errors from different classifiers (e.g., [78] and [112]). Various classifier fusion methods (fixed rules and trained combiners) have been proposed in the literature, but optimal design methods for classifier fusion are still not available. In addition, since humans simultaneously employ the tightly coupled audio and visual modalities, the multimodal signals cannot be considered mutually independent and should not be combined only in the end, as in the case of decision-level fusion.

Model-level fusion or hybrid fusion that aims at combining the benefits of both feature-level and decision-level fusion methods may be a good choice for this fusion problem. However, based on existing knowledge and methods, how one can model multimodal fusion based on multilabel multi-time-scale labeling scheme mentioned above is largely unexplored. A number of issues relevant to fusion require further investigation, such as the optimal level of integrating these different streams, the optimal function for the integration, and the inclusion of suitable estimations of the reliability of each stream. In addition, how one can build context-dependent multimodal fusion is an open and highly relevant issue.

Here, we want to stress that temporal structures of the modalities (facial and vocal) and their temporal correlations play an extremely important role in the interpretation of human naturalistic audiovisual affective behavior (see Section 2 for a discussion). Yet, these are virtually unexplored areas of research due to the fact that the facial expression and vocal expression of emotion are usually studied separately.

4.5 A Few Additional Related Issues

Context. An important related issue that should be addressed in all visual, vocal, and audiovisual affect recognition is how one can make use of information about the context (environment, observed subject, or the current task), in which the observed affective behavior was displayed. Affects are intimately related to a situation being experienced or imagined by humans. Without context, humans may misunderstand the observed person's emotion expressions. Yet, with the exception of a few studies that investigated the influence of context on affect recognition (e.g., [50], [52], [69], [72], [86], and [104]), virtually all existing approaches to the machine analysis of human affect are context insensitive. Building a context model that includes person ID, gender, age, conversation topic, and workload need help from other research fields like face recognition, gender recognition, age recognition, topic detection, and task tracking. Since the problem of context sensing is very difficult to solve, pragmatic approaches (e.g., activity and/or subject-profiled approaches) should be taken.

Segmentation. Almost all of the existing methods are tested just on presegmented emotion sequences or images, except for a few studies (e.g., [11] and [25]) that use heuristic methods for segmenting the emotions from videos. Automatic continuous emotion recognition is a dynamic searching process that continuously makes emotion inference in the presence of signal ambiguity and context. This is rather complicated, since the search algorithm has to consider the possibility of each emotion starting at any arbitrary time frame. Furthermore, the number of emotions changing in a video is not known, and the boundaries between different emotional expressions are full of ambiguity. It becomes more challenging in multimodal affect recognition because different modalities (e.g., face, body, and vocal expressions) have different temporal structures and often do not synchronize. If we aim at developing a practical affect recognizer, the emotion segmentation is one of the most important issues but has not been largely unexplored so far.

Evaluation. The existing methods for the machine analysis of human affect surveyed and discussed throughout this paper are difficult to compare because they are rarely (if ever) tested on a common data set. United efforts of the relevant research communities are needed to specify evaluation procedures that could be used for establishing reliable measures of systems' performance based on a comprehensive readily accessible benchmark database.

5 Conclusion

Research on the machine analysis of human affect has witnessed a good deal of progress when compared to that described in the survey papers of Pantic and Rothkrantz in 2003 [102] and Cowie et al. in 2001 [31]. At those times, a few small-sized data sets of affective displays existed, and almost all methods for the machine analysis of human affect were unimodal based on deliberate displays of either facial expressions or vocal expressions of six prototypical emotions. Available data was not shared among researchers, multimedia data and multimodal human affect analyzers were rare, and the machine analysis of spontaneous displays of affective behavior seemed to be in a distant future. Today, several large collections of acted affective displays are shared by the researchers in the field, and some data sets of spontaneously displayed expressions have been recently made available. A number of promising methods for vision-based, audio-based, and audiovisual analysis of human spontaneous behavior have been proposed. This paper focused on surveying and discussing these novel approaches to the machine analysis of human affect and on summarizing the issues that have not received sufficient attention but are crucial for advancing the machine interpretation of human behavior in naturalistic contexts. The most important of these issues yet to be addressed in the field include the following:

- building a comprehensive readily accessible reference set of affective displays, which could provide a basis for benchmarks for all different efforts in the research on the machine analysis of human affective behavior, and defining the appropriate evaluation procedures,
- developing methods for spontaneous affective behavior analysis, which are robust to observed person's arbitrary movement, occlusion, and complex and noisy background,
- devising models and methods for human affect analysis, which take into consideration the temporal structures of the modalities and the temporal correlations between the modalities (and/or multiple cues) and context (subject, the task, and environment), and
- developing better methods for multimodal fusion.

Since the complexity of these issues concerned with the interpretation of human behavior at a deeper level is tremendous and spans several different disciplines in computer and social sciences, we believe that a large interdisciplinary international program directed toward computer understanding of human behavioral patterns should be established if we are to experience true breakthroughs in this and the related research fields. The progress in research on the machine analysis of human affect can aid in the creation of a new paradigm for HCI (affect-sensitive interfaces and socially intelligent environments) and advance the research in several related fields, including psychology, psychiatry, and education.

Acknowledgments

The authors would like to thank Qiang Ji and the anonymous reviewers for encouragement and valuable comments. This paper is a collaborative work. Thomas Huang is the leader of this team work but prefers to be the last in the author list. Zhihong Zeng wrote the first draft, Maja Pantic significantly improved it by rewriting it and offering important advice, and Glenn Roisman provided important comments and polished the whole paper. Zhihong Zeng and Thomas S. Huang's work in this paper was supported in part by a Beckman Postdoctoral Fellowship, US National Science Foundation Grant CCF 04-26627 and the US Government VACE Program. Maja Pantic's research that lead to these results was funded in part by the EC FP7 Programme [FP7/2007-2013] under grant agreement no. 211486 (SEMAINE) and the European Research Council under the ERC Starting Grant agreement No. ERC-2007-StG-203143 (MAHNOB).

References

[1] N. Ambady and R. Rosenthal, "Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences," *A Meta-Analysis: Psychological Bull.,* vol. 111, no. 2, pp. 256-274, 1992.

[2] K. Anderson and P.W. McOwan, "A Real-Time Automated System for Recognition of Human Facial Expressions," *IEEE Trans. Systems, Man, and Cybernetics Part B,* vol. 36, no. 1, pp. 96-105, 2006.

[3] J. Ang et al., "Prosody-Based Automatic Detection of Annoyance and Frustration in Human-Computer Dialog," *Proc. Eighth Int'l Conf. Spoken Language Processing (ICSLP),* 2002.

[4] A.B. Ashraf, S. Lucey, J.F. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, and B.J. Theobald, "The Painful Face: Pain Expression Recognition Using Active Appearance Models," *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07),* pp. 9-14, 2007.

[5] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for Emotional Speech: Clarifying the Issues and Enhancing Performance," *Neural Networks,* vol. 18, pp. 437-444, 2005.

[6] A. Austermann, N. Esau, L. Kleinjohann, and B. Kleinjohann, "Prosody-Based Emotion Recognition for MEXI," *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS '05),* pp. 1138-1144, 2005.

[7] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion Analysis in Man-Machine Interaction Systems," *LNCS 3361,* pp. 318-328, 2005.

[8] R. Banse and K.R. Scherer, "Acoustic Profiles in Vocal Emotion Expression," *J. Personality Social Psychology,* vol. 70, no. 3, pp. 614-636, 1996.

[9] M.S. Bartlett, G. Littlewort, P. Braathen, T.J. Sejnowski, and J.R. Movellan, "A Prototype for Automatic Recognition of Spontaneous Facial Actions," *Advances in Neural Information Processing Systems,* vol. 15, pp. 1271-1278, 2003.

[10] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '05),* pp. 568-573, 2005.

[11] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Fully Automatic Facial Action Recognition in Spontaneous Behavior," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '06),* pp. 223-230, 2006.

[12] A. Batliner, K. Fischer, R. Hubera, J. Spilkera, and E. Noth, "How to Find Trouble in Communication," *Speech Comm.,* vol. 40, pp. 117-143, 2003.

[13] A. Batliner et al., "You Stupid Tin Box—Children Interacting with the AIBO Robot: A Cross-Linguistic Emotional Speech," *Proc. Fourth Int'l Conf. Language Resources and Evaluation,* 2004.

[14] C. Blouin and V. Maffiolo, "A Study on the Automatic Detection and Characterization of Emotion in a Voice Service Context," *Proc. Ninth European Conf. Speech Comm. and Technology (INTERSPEECH '05),* pp. 469-472, 2005.

[15] S. Burger, V. MacLaren, and H. Yu, "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style," *Proc. Eighth Int'l Conf. Spoken Language Processing (ICSLP),* 2002.

[16] C. Busso et al., "Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information," *Proc. Sixth ACM Int'l Conf. Multimodal Interfaces (ICMI '04),* pp. 205-211, 2004.

[17] G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Paouzaiou, and K. Karpouzis, "Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition," *Proc. Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06),* pp. 146-154, 2006.

[18] Y. Chang, C. Hu, and M. Turk, "Probabilistic Expression Analysis on Manifolds," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '04),* vol. 2, pp. 520-527, 2004.

[19] Y. Chang, C. Hu, R. Feris, and M. Turk, "Manifold Based Analysis of Facial Expression," *J. Image and Vision Computing,* vol. 24, no. 6, pp. 605-614, 2006.

[20] Y. Chang, M. Vieira, M. Turk, and L. Velho, "Automatic 3D Facial Expression Analysis in Videos," *Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures (AMFG '05),* vol. 3723, pp. 293-307, 2005.

[21] L.S. Chen, "Joint Processing of Audio-Visual Information for the Recognition of Emotional Expressions in Human-Computer Interaction," PhD dissertation, Univ. of Illinois, Urbana-Champaign, 2000.

[22] L. Chen, T.S. Huang, T. Miyasato, and R. Nakatsu, "Multimodal Human Emotion/Expression Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '98),* pp. 396-401, 1998.

[23] J.A. Coan and J.J.B. Allen, *Handbook of Emotion Elicitation and Assessment.* Oxford Univ. Press, 2007.

[24] I. Cohen, F. Cozman, N. Sebe, M. Cirelo, and T.S. Huang, "Semi-Supervised Learning of Classifiers: Theory, Algorithms, and Their Applications to Human-Computer Interaction," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 12, pp. 1553-1567, Dec. 2004.

[25] L. Cohen, N. Sebe, A. Garg, L. Chen, and T. Huang, "Facial Expression Recognition from Video Sequences: Temporal and Static Modeling," *Computer Vision and Image Understanding,* vol. 91, nos. 1-2, pp. 160-187, 2003.

[26] J.F. Cohn, "Foundations of Human Computing: Facial Expression and Emotion," *Proc. Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06),* pp. 233-238, 2006.

[27] J.F. Cohn, L.I. Reed, Z. Ambadar, J. Xiao, and T. Moriyama, "Automatic Analysis and Recognition of Brow Actions and Head Motion in Spontaneous Facial Behavior," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '04),* vol. 1, pp. 610-616, 2004.

[28] J.F. Cohn and K.L. Schmidt, "The Timing of Facial Motion in Posed and Spontaneous Smiles," *Int'l J. Wavelets, Multiresolution and Information Processing,* vol. 2, pp. 1-12, 2004.

[29] J.F. Cohn and E.Z. Tronick, "Mother-Infant Interaction: The Sequence of Dyadic States at Three, Six, and Nine Months," *Development Psychology,* vol. 23, pp. 68-77, 1988.

[30] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'Feeltrace': An Instrument for Recording Perceived Emotion in Real Time," *Proc. ISCA Workshop Speech and Emotion,* pp. 19-24, 2000.

[31] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine,* pp. 32-80, Jan. 2001.

[32] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond Emotion Archetypes: Databases for Emotion Modeling Using Neural Networks," *Neural Networks,* vol. 18, pp. 371-388, 2005.

[33] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing Emotion in Speech," *Proc. Fourth Int'l Conf. Spoken Language Processing (ICSLP '96),* pp. 1970-1973, 1996.

[34] L. Devillers and I. Vasilescu, "Reliability of Lexical and Prosodic Cues in Two Real-Life Spoken Dialog Corpora," *Proc. Fourth Int'l Conf. Language Resources and Evaluation (LREC),* 2004.

[35] L. Devillers and I. Vasilescu, "Real-Life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs," *Proc. Ninth Int'l Conf. Spoken Language Processing (ICSLP),* 2006.

[36] I. Vasilescu and L. Devillers, "Detection of Real-Life Emotions in Call Centers," *Proc. Ninth European Conf. Speech Comm. and Technology (INTERSPEECH),* 2005.

[37] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in Real-Life Emotion Annotation and Machine Learning Based Detection," *Neural Networks,* vol. 18, pp. 407-422, 2005.

[38] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional Speech: Towards a New Generation of Database," *Speech Comm.,* vol. 40, nos. 1/2, pp. 33-60, 2003.

[39] Z. Duric, W.D. Gray, R. Heishman, F. Li, A. Rosenfeld, M.J. Schoelles, C. Schunn, and H. Wechsler, "Integrating Perceptual and Cognitive Modeling for Adaptive and Intelligent Human-Computer Interaction," *Proc. IEEE,* vol. 90, no. 7, pp. 1272-1289, 2002.

[40] P. Ekman, "Universals and Cultural Differences in Facial Expressions of Emotion," *Proc. Nebraska Symp. Motivation,* pp. 207-283, 1971.

[41] *Emotion in the Human Face,* P. Ekman, ed., second ed. Cambridge Univ. Press, 1982.

[42] P. Ekman, "Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique," *Psychological Bull.,* vol. 115, no. 2, pp. 268-287, 1994.

[43] P. Ekman, W.V. Friesen, and J.C. Hager, "Facial Action Coding System," *A Human Face,* 2002.

[44] *NSF Understanding the Face: A Human Face eStore,* P. Ekman, T.S. Huang, T.J. Sejnowski, and J.C. Hager, eds., 1993.

[45] P. Ekman, D. Matsumoto, and W.V. Friesen, "Facial Expression in Affective Disorders," *What the Face Reveals,* P. Ekman and E.L. Rosenberg, eds., pp. 429-439, 2005.

[46] P. Ekman and H. Oster, "Facial Expressions of Emotion," *Ann. Rev. Psychology,* vol. 30, pp. 527-554, 1979.

[47] P. Ekman and E.L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System,* second ed. Oxford Univ. Press, 2005.

[48] R. El Kaliouby and P. Robinson, "Real-Time Inference of Complex Mental States from Facial Expression and Head Gestures," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '04),* vol. 3, p. 154, 2004.

[49] B. Fasel and J. Luttin, "Automatic Facial Expression Analysis: Survey," *Pattern Recognition,* vol. 36, no. 1, pp. 259-275, 2003.

[50] B. Fasel, F. Monay, and D. Gatica-Perez, "Latent Semantic Analysis of Facial Action Codes for Automatic Facial Expression Recognition," *Proc. Sixth ACM Int'l Workshop Multimedia Information Retrieval (MIR '04),* pp. 181-188, 2004.

[51] C.-N. Fiechter, "Efficient Reinforcement Learning," *Proc. Seventh Ann. ACM Conf. Computational Learning Theory,* pp. 88-97, 1994.

[52] K. Forbes-Riley and D. Litman, "Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources," *Proc. Human Language Technology Conf. North Am. Chapter of the Assoc. Computational Linguistics (HLT/NAACL),* 2004.

[53] F. Fragopanagos and J.G. Taylor, "Emotion Recognition in Human-Computer Interaction," *Neural Networks,* vol. 18, pp. 389-405, 2005.

[54] E. Fried, "The Impact of Nonverbal Communication of Facial Affect on Children's Learning," PhD dissertation, Rutgers Univ., 1976.

[55] G. Furnas, T. Landauer, L. Gomes, and S. Dumais, "The Vocabulary Problem in Human-System Communication," *Comm. ACM,* vol. 30, no. 11, pp. 964-972, 1987.

[56] H.J. Go, K.C. Kwak, D.J. Lee, and M.G. Chun, "Emotion Recognition from Facial Image and Speech Signal," *Proc. Int'l Conf. Soc. of Instrument and Control Engineers,* pp. 2890-2895, 2003.

[57] M. Graciarena, E. Shriberg, A. Stolcke, F. Enos, J. Hirschberg, and S. Kajarekar, "Combining Prosodic, Lexical and Cepstral Systems for Deceptive Speech Detection," *Proc. Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP '06),* vol. I, pp. 1033-1036, 2006.

[58] M. Greenwald, E. Cook, and P. Lang, "Affective Judgment and Psychophysiological Response: Dimensional Covariation in the Evaluation of Pictorial Stimuli," *J. Psychophysiology,* vol. 3, pp. 51-64, 1989.

[59] R. Gross, "Face Databases," *Handbook of Face Recognition,* S.Z. Li and A.K. Jain, eds., pp. 301-328, Springer, 2005.

[60] H. Gu and Q. Ji, "An Automated Face Reader for Fatigue Detection," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '04),* pp. 111-116, 2004.

[61] H. Gunes and M. Piccardi, "Affect Recognition from Face and Body: Early Fusion versus Late Fusion," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '05),* pp. 3437-3443, 2005.

[62] H. Gunes and M. Piccardi, "Fusing Face and Body Display for Bi-Modal Emotion Recognition: Single Frame Analysis and Multi-Frame Post-Integration," *Proc. First Int'l Conf. Affective Computing and Intelligent Interaction (ACII '05),* pp. 102-111, 2005.

[63] H. Gunes and M. Piccardi, "A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior," *Proc. 18th Int'l Conf. Pattern Recognition (ICPR '06),* vol. 1, pp. 1148-1153, 2006.

[64] G. Guo and C.R. Dyer, "Learning from Examples in the Small Sample Case: Face Expression Recognition," *IEEE Trans. Systems, Man, and Cybernetics Part B,* vol. 35, no. 3, pp. 477-488, 2005.

[65] J. Hirschberg, S. Benus, J.M. Brenier, F. Enos, and S. Friedman, "Distinguishing Deceptive from Non-Deceptive Speech," *Proc. Ninth European Conf. Speech Comm. and Technology (INTERSPEECH '05),* pp. 1833-1836, 2005.

[66] S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll, "Bimodal Fusion of Emotional Data in an Automotive Environment," *Proc. 30th Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05),* vol. II, pp. 1085-1088, 2005.

[67] S. Ioannou, A. Raouzaiou, V. Tzouvaras, T. Mailis, K. Karpouzis, and S. Kollias, "Emotion Recognition through Facial Expression Analysis Based on a Neurofuzzy Method," *Neural Networks,* vol. 18, pp. 423-435, 2005.

[68] A. Jaimes and N. Sebe, "Multimodal Human Computer Interaction: A Survey," *Proc. 11th IEEE Int'l Workshop Human Computer Interaction (HCI),* 2005.

[69] Q. Ji, P. Lan, and C. Looney, "A Probabilistic Framework for Modeling and Real-Time Monitoring Human Fatigue," *IEEE Systems, Man, and Cybernetics Part A,* vol. 36, no. 5, pp. 862-875, 2006.

[70] P.N. Juslin and K.R. Scherer, "Vocal Expression of Affect," *The New Handbook of Methods in Nonverbal Behavior Research,* J. Harrigan, R. Rosenthal, and K. Scherer, eds., Oxford Univ. Press, 2005.

[71] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Face and Gesture Recognition (AFGR '00),* pp. 46-53, 2000.

[72] A. Kapoor, W. Burleson, and R.W. Picard, "Automatic Prediction of Frustration," *Int'l J. Human-Computer Studies,* vol. 65, no. 8, pp. 724-736, 2007.

[73] A. Kapoor and R.W. Picard, "Multimodal Affect Recognition in Learning Environment," *Proc. 13th ACM Int'l Conf. Multimedia (Multimedia '05),* pp. 677-682, 2005.

[74] K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaiou, L. Malatesta, and S. Kollias, "Modeling Naturalistic Affective States via Facial, Vocal, and Bodily Expression Recognition," *LNAI 4451,* pp. 91-112, 2007.

[75] D. Keltner, "Signs of Appeasement: Evidence for the Distinct Displays of Embarrassment, Amusement and Shame," *J. Personality and Social Psychology,* vol. 68, no. 3, pp. 441-454, 1995.

[76] H. Kobayashi and F. Hara, "The Recognition of Basic Facial Expressions by Neural Network," *Proc. IEEE Int'l Joint Conf. Neural Networks (IJCNN '91),* pp. 460-466, 1991.

[77] I. Kotsia and I. Pitas, "Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines," *IEEE Trans. Image Processing,* vol. 16, no. 1, pp. 172-187, 2007.

[78] L.I. Kuncheva, *Combining Pattern Classifier: Methods and Algorithms.* John Wiley & Sons, 2004.

[79] O.W. Kwon, K. Chan, J. Hao, and T.W. Lee, "Emotion Recognition by Speech Signals," *Proc. Eighth European Conf. Speech Comm. and Technology (EUROSPEECH),* 2003.

[80] K. Laskowski and S. Burger, "Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus," *Proc. Fifth Int'l Conf. Language Resources and Evaluation (LREC),* 2006.

[81] C. Lee and A. Elgammal, "Facial Expression Analysis Using Nonlinear Decomposable Generative Models," *Proc. Second IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures (AMFG),* 2005.

[82] C. Lee and S. Narayanan, "Emotion Recognition Using a Data-Driven Fuzzy Inference System," *Proc. Eighth European Conf. Speech Comm. and Technology (EUROSPEECH '03),* pp. 157-160, 2003.

[83] C.M. Lee and S.S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," *IEEE Trans. Speech and Audio Processing,* vol. 13, no. 2, pp. 293-303, 2005.

[84] J. Liscombe, J. Hirschberg, and J.J. Venditti, "Detecting Certainness in Spoken Tutorial Dialogues," *Proc. Ninth European Conf. Speech Comm. and Technology (INTERSPEECH),* 2005.

[85] C.L. Lisetti and F. Nasoz, "MAUI: A Multimodal Affective User Interface," *Proc. 10th ACM Int'l Conf. Multimedia (Multimedia '02),* pp. 161-170, 2002.

[86] D.J. Litman and K. Forbes-Riley, "Predicting Student Emotions in Computer-Human Tutoring Dialogues," *Proc. 42nd Ann. Meeting of the Assoc. Computational Linguistics (ACL '04),* July 2004.

[87] G.C. Littlewort, M.S. Bartlett, and K. Lee, "Faces of Pain: Automated Measurement of Spontaneous Facial Expressions of Genuine and Posed Pain," *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07),* pp. 15-21, 2007.

[88] S. Lucey, A.B. Ashraf, and J.F. Cohn, "Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face," *Face Recognition,* K. Delac, and M. Grgic, eds., pp. 275-286, I-Tech Education and Publishing, 2007.

[89] L. Maat and M. Pantic, "Gaze-X: Adaptive Affective Multimodal Interface for Single-User Office Scenarios," *Proc. Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06),* pp. 171-178, 2006.

[90] K. Mase, "Recognition of Facial Expression from Optical Flow," *IEICE Trans.,* vol. 74, no. 10, pp. 3474-3483, 1991.

[91] S. Matos, S.S. Birring, I.D. Pavord, and D.H. Evans, "Detection of Cough Signals in Continuous Audio Recordings Using HMM," *IEEE Trans. Biomedical Eng.,* vol. 53, no. 6, pp. 1078-1083, 2006.

[92] A. Mehrabian, "Communication with Words," *Psychology Today,* vol. 2, no. 4, pp. 53-56, 1968.

[93] S. Mozziconacci, "Prosody and Emotions," *Proc. First Int'l Conf. Speech Prosody (Speech Prosody '02),* pp. 1-9, 2002.

[94] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion Recognition in Spontaneous Speech Using GMM," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '06),* pp. 809-812, 2006.

[95] A.J. O'Toole et al., "A Video Database of Moving Faces and People," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 812-816, May 2005.

[96] P.-Y. Oudeyer, "The Production and Recognition of Emotions in Speech: Features and Algorithms," *Int'l J. Human-Computer Studies,* vol. 59, pp. 157-183, 2003.

[97] P. Pal, A.N. Iyer, and R.E. Yantorno, "Emotion Detection from Infant Facial Expressions and Cries," *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP '06),* vol. 2, pp. 721-724, 2006.

[98] M. Pantic and M.S. Bartlett, "Machine Analysis of Facial Expressions," *Face Recognition,* K. Delac and M. Grgic, eds., pp. 377-416, I-Tech Education and Publishing, 2007.

[99] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments Form Face Profile Image Sequences," *IEEE Trans. Systems, Man, and Cybernetics Part B,* vol. 36, no. 2, pp. 433-449, 2006.

[100] M. Pantic, A. Pentland, A. Nijholt, and T.S. Huang, "Human Computing and Machine Understanding of Human Behavior: A Survey," *Proc. Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06),* pp. 239-248, 2006.

[101] M. Pantic and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[102] M. Pantic and L.J.M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human-Computer Interaction," *Proc. IEEE,* vol. 91, no. 9, pp. 1370-1390, Sept. 2003.

[103] M. Pantic and L.J.M. Rothkrantz, "Facial Action Recognition for Facial Expression Analysis from Static Face Images," *IEEE Trans. Systems, Man, and Cybernetics Part B,* vol. 34, no. 3, pp. 1449-1461, 2004.

[104] M. Pantic and L.J.M. Rothkrantz, "Case-Based Reasoning for User-Profiled Recognition of Emotions from Face Images," *Proc. 12th ACM Int'l Conf. Multimedia (Multimedia '04),* pp. 391-394, 2004.

[105] M. Pantic, N. Sebe, J.F. Cohn, and T. Huang, "Affective Multimodal Human-Computer Interaction," *Proc. 13th ACM Int'l Conf. Multimedia (Multimedia '05),* pp. 669-676, 2005.

[106] M. Pantic, M.F. Valstar, R. Rademaker, and L. Maat, "Web-Based Database for Facial Expression Analysis," *Proc. 13th ACM Int'l Conf. Multimedia (Multimedia '05),* pp. 317-321, 2005.

[107] A. Pentland, "Socially Aware, Computation and Communication," *Computer,* vol. 38, pp. 33-40, 2005.

[108] S. Petridis and M. Pantic, "Audiovisual Discrimination between Laughter and Speech," *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP),* pp. 5117-5120, 2008.

[109] R.W. Picard, *Affective Computing.* MIT Press, 1997.

[110] R. Plutchik, *Emotion: A Psychoevolutionary Synthesis.* Harper and Row, 1980.

[111] G.I. Roisman, J.L. Tsai, and K.S. Chiang, "The Emotional Integration of Childhood Experience: Physiological, Facial Expressive, and Self-Reported Emotional Response during the Adult Attachment Interview," *Developmental Psychology,* vol. 40, no. 5, pp. 776-789, 2004.

[112] *Proc. Int'l Workshop Multiple Classifier Systems (MCS),* F. Roli et al., eds., 2001-2005.

[113] J.A. Russell, J. Bachorowski, and J. Fernandez-Dols, "Facial and Vocal Expressions of Emotion," *Ann. Rev. Psychology,* vol. 54, pp. 329-349, 2003.

[114] J. Russell and A. Mehrabian, "Evidence for a Three-Factor Theory of Emotions," *J. Research in Personality,* vol. 11, pp. 273-294, 1977.

[115] A. Samal and P.A. Iyengar, "Automatic Recognition and Analysis of Human Faces and Facial Expressions: A Survey," *Pattern Recognition,* vol. 25, no. 1, pp. 65-77, 1992.

[116] D. Sander, D. Grandjean, and K.R. Scherer, "A System Approach to Appraisal Mechanisms in Emotion," *Neural Networks,* vol. 18, pp. 317-352, 2005.

[117] K.R. Scherer, "Appraisal Theory," *Handbook of Cognition and Emotion,* T. Dalgleish and M.J. Power, eds., pp. 637-663, Wiley, 1999.

[118] B. Schuller, R. Muller, B. Hornler, A. Hothker, H. Konosu, and G. Rigoll, "Audiovisual Recognition of Spontaneous Interest within Conversations," *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07),* pp. 30-37, 2007.

[119] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-Robust Speech Recognition by Dynamic Emotional Adaptation," *Proc. Speech Prosody,* Special Session on Prosody in Automatic Speech Recognition, 2006.

[120] B. Schuller, R.J. Villar, G. Rigoll, and M. Lang, "Meta-Classifiers in Acoustic and Linguistic Feature Fusion-Based Affect Recognition," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05),* pp. 325-328, 2005.

[121] N. Sebe, I. Cohen, and T.S. Huang, "Multimodal Emotion Recognition," *Handbook of Pattern Recognition and Computer Vision,* World Scientific, 2005.

[122] N. Sebe, I. Cohen, T. Gevers, and T.S. Huang, "Emotion Recognition Based on Joint Visual and Audio Cues," *Proc. 18th Int'l Conf. Pattern Recognition (ICPR '06),* pp. 1136-1139, 2006.

[123] N. Sebe, M.S. Lew, I. Cohen, Y. Sun, T. Gevers, and T.S. Huang, "Authentic Facial Expression Analysis," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR),* 2004.

[124] M. Song, J. Bu, C. Chen, and N. Li, "Audio-Visual-Based Emotion Recognition: A New Approach," *Proc. Int'l Conf. Computer Vision and Pattern Recognition (CVPR '04),* pp. 1020-1025, 2004.

[125] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, "Off All Things the Measure Is Man: Automatic Classification of Emotions and Inter-Labeler Consistency," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05),* vol. 1, pp. 317-320, 2005.

[126] B. Stein and M.A. Meredith, *The Merging of Senses.* MIT Press, 1993.

[127] M. Suwa, N. Sugie, and K. Fujimora, "A Preliminary Note on Pattern Recognition of Human Emotional Expression," *Proc. Int'l Joint Conf. Pattern Recognition,* pp. 408-410, 1978.

[128] H. Tao and T.S. Huang, "Explanation-Based Facial Motion Tracking Using a Piecewise Bezier Volume Deformation Mode," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '99),* vol. 1, pp. 611-617, 1999.

[129] A. Teeters, R.E. Kaliouby, and R.W. Picard, "Self-Cam: Feedback from What Would Be Your Social Partner," *Proc. ACM SIGGRAPH '06,* p. 138, 2006.

[130] Y.L. Tian, T. Kanade, and J.F. Cohn, "Facial Expression Analysis," *Handbook of Face Recognition,* S.Z. Li and A.K. Jain, eds., pp. 247-276, Springer, 2005.

[131] S.S. Tomkins, *Affect, Imagery, Consciousness,* vol. 1. Springer, 1962.

[132] Y. Tong, W. Liao, and Q. Ji, "Facial Action Unit Recognition by Exploiting Their Dynamics and Semantic Relationships," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 29, no. 10, pp. 1683-1699, 2007.

[133] K.P. Truong and D.A. van Leeuwen, "Automatic Discrimination between Laughter and Speech," *Speech Comm.,* vol. 49, pp. 144-158, 2007.

[134] M.F. Valstar, H. Gunes, and M. Pantic, "How to Distinguish Posed from Spontaneous Smiles Using Geometric Features," *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07),* pp. 38-45, 2007.

[135] M. Valstar, M. Pantic, Z. Ambadar, and J.F. Cohn, "Spontaneous versus Posed Facial Behavior: Automatic Analysis of Brow Actions," *Proc. Eight Int'l Conf. Multimodal Interfaces (ICMI '06),* pp. 162-170, 2006.

[136] M. Valstar, M. Pantic, and I. Patras, "Motion History for Facial Action Detection from Face Video," *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics (SMC '04),* vol. 1, pp. 635-640, 2004.

[137] H. Wang and N. Ahuja, "Facial Expression Decomposition," *Proc. Ninth IEEE Int'l Conf. Computer Vision (ICCV '03),* p. 958, 2003.

[138] Y. Wang and L. Guan, "Recognizing Human Emotion from Audiovisual Information," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05),* pp. 1125-1128, 2005.

[139] J. Wang, L. Yin, X. Wei, and Y. Sun, "3D Facial Expression Recognition Based on Primitive Surface Feature Distribution," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '06),* vol. 2, pp. 1399-1406, 2006.

[140] D. Watson, L.A. Clark, and A. Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales," *J. Personality and Social Psychology,* vol. 54, pp. 1063-1070, 1988.

[141] Z. Wen and T.S. Huang, "Capturing Subtle Facial Motions in 3D Face Tracking," *Proc. Ninth IEEE Int'l Conf. Computer Vision (ICCV '03),* pp. 1343-1350, 2003.

[142] C.M. Whissell, "The Dictionary of Affect in Language," *Emotion: Theory, Research and Experience. The Measurement of Emotions,* R. Plutchik and H. Kellerman, eds., vol. 4, pp. 113-131, Academic Press, 1989.

[143] J. Whitehill and C.W. Omlin, "Haar Features for FACS AU Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '06),* pp. 217-222, 2006.

[144] A.C. de C. Williams, "Facial Expression of Pain: An Evolutionary Account," *Behavioral and Brain Sciences,* vol. 25, no. 4, pp. 439-488, 2002.

[145] C. Williams and K. Stevens, "Emotions and Speech: Some Acoustic Correlates," *J. Acoustic Soc. of Am.,* vol. 52, no. 4, pp. 1238-1250, 1972.

[146] J. Xiao, T. Moriyama, T. Kanade, and J.F. Cohn, "Robust Full-Motion Recovery of Head by Dynamic Templates and Re-Registration Techniques," *Int'l J. Imaging Systems and Technology,* vol. 13, no. 1, pp. 85-94, 2003.

[147] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of Facial Expressions and Measurement of Levels of Interest from Video," *IEEE Trans. Multimedia,* vol. 8, no. 3, pp. 500-507, June 2006.

[148] L. Yin, X. Wei, Y. Sun, J. Wang, and M.J. Rosato, "A 3D Facial Expression Database for Facial Behavior Research," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition (AFGR '06),* pp. 211-216, 2006.

[149] Z. Zeng, Y. Fu, G.I. Roisman, Z. Wen, Y. Hu, and T.S. Huang, "Spontaneous Emotional Facial Expression Detection," *J. Multimedia,* vol. 1, no. 5, pp. 1-8, 2006.

[150] Z. Zeng, Y. Hu, M. Liu, Y. Fu, and T.S. Huang, "Training Combination Strategy of Multi-Stream Fused Hidden Markov Model for Audio-Visual Affect Recognition," *Proc. 14th ACM Int'l Conf. Multimedia (Multimedia '06),* pp. 65-68, 2006.

[151] Z. Zeng, Y. Hu, G.I. Roisman, Z. Wen, Y. Fu, and T.S. Huang, "Audio-Visual Spontaneous Emotion Recognition," *Artificial Intelligence for Human Computing,* T.S. Huang, A. Nijholt, M. Pantic, and A. Pentland, eds., pp. 72-90, Springer, 2007.

[152] Z. Zeng, M. Pantic, G.I. Roisman, and T.S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces (ICMI '07),* pp. 126-133, 2007.

[153] Z. Zeng, J. Tu, M. Liu, T. Zhang, N. Rizzolo, Z. Zhang, T.S. Huang, D. Roth, and S. Levinson, "Bimodal HCI-Related Emotion Recognition," *Proc. Sixth ACM Int'l Conf. Multimodal Interfaces (ICMI '04),* pp. 137-143, 2004.

[154] Z. Zeng, J. Tu, P. Pianfetti, M. Liu, T. Zhang, Z. Zhang, T.S. Huang, and S. Levinson, "Audio-Visual Affect Recognition through Multi-Stream Fused HMM for HCI," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '05),* pp. 967-972, 2005.

[155] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-Visual Affect Recognition," *IEEE Trans. Multimedia,* vol. 9, no. 2, pp. 424-428, Feb. 2007.

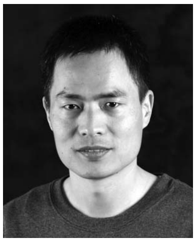
[156] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T.S. Huang, "Audio-Visual Affect Recognition in Activation-Evaluation Space," *Proc. 13th ACM Int'l Conf. Multimedia (Multimedia '05),* pp. 828-831, 2005.

[157] T. Zhang, M. Hasegawa-Johnson, and S.E. Levinson, "Children's Emotion Recognition in an Intelligent Tutoring Scenario," *Proc. Eighth European Conf. Speech Comm. and Technology (INTERSPEECH),* 2004.

[158] Y. Zhang and Q. Ji, "Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 699-714, 2005.

[159] Z.-H. Zhou, K.-J. Chen, and H.-B. Dai, "Enhancing Relevance Feedback in Image Retrieval Using Unlabeled Data," *ACM Trans. Information Systems,* vol. 24, no. 2, pp. 219-244, 2006.

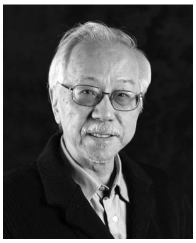
[160] Z. Zhu and Q. Ji, "Robust Real-Time Face Pose and Facial Expression Recovery," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR '06),* vol. 1, pp. 681-688, 2006.



Zhihong Zeng received the PhD degree from the Chinese Academy of Sciences in 2002. He is currently a Beckman postdoctoral fellow at the Beckman Institute, University of Illinois, Urbana-Champaign (UIUC). His research interests include multimodal affective computing, multimodal human-computer interaction, and computer vision. He is a member of the IEEE Computer Society.

Maja Pantic received the MSc and PhD degrees in computer science from the Delft University of Technology, The Netherlands, in 1997 and 2001, respectively. She is currently a reader in Multimodal HCI in the Department of Computing, Imperial College London, and a professor of affective and behavioral computing in the Department of Computer Science, University of Twente. She is an associate editor for the *IEEE Transactions on Systems, Man, and Cybernetics Part B* and the *Image and Vision Computing Journal*. She is a guest editor, organizer, and committee member of more than 10 major journals and conferences. Her research interests include computer vision and machine learning applied to face and body gesture recognition, multimodal human-computer interaction (HCI), context-sensitive HCI, affective computing, and e-learning tools. She is a senior member of the IEEE.

Glenn I. Roisman received the PhD degree from the University of Minnesota in 2002. He is currently an assistant professor in the Department of Psychology, University of Illinois, Urbana-Champaign (UIUC). His research interests include social and emotional development across the lifespan. He has published more than 25 scholarly journal articles and book chapters. He received the Award for Early Research Contributions from the Society for Research in Child Development in 2007.



Thomas S. Huang received the DSc degree in electrical engineering from the Massachusetts Institute of Technology (MIT). He has been with the faculty of MIT and Purdue University. In 1980, he joined the University of Illinois, Urbana-Champaign, where he is currently the William L. Everitt Distinguished Professor of Electrical and Computer Engineering, a research professor in the Coordinated Science Laboratory, and a cochair of the Human Computer Intelligent Interactive Major Research Team, Beckman Institute. He has published 21 books and more than 600 papers in network theory, tpdedigital holograph, image and video compression, multimodal human computer interfaces, and multimedia databases. He is a fellow of the IEEE and a member of the National Academy of Engineering. He has received many honors and awards, including the IEEE Jack S. Kilby Signal Processing Medal and the King-Sun Fu Prize from the International Association of Pattern Recognition.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.