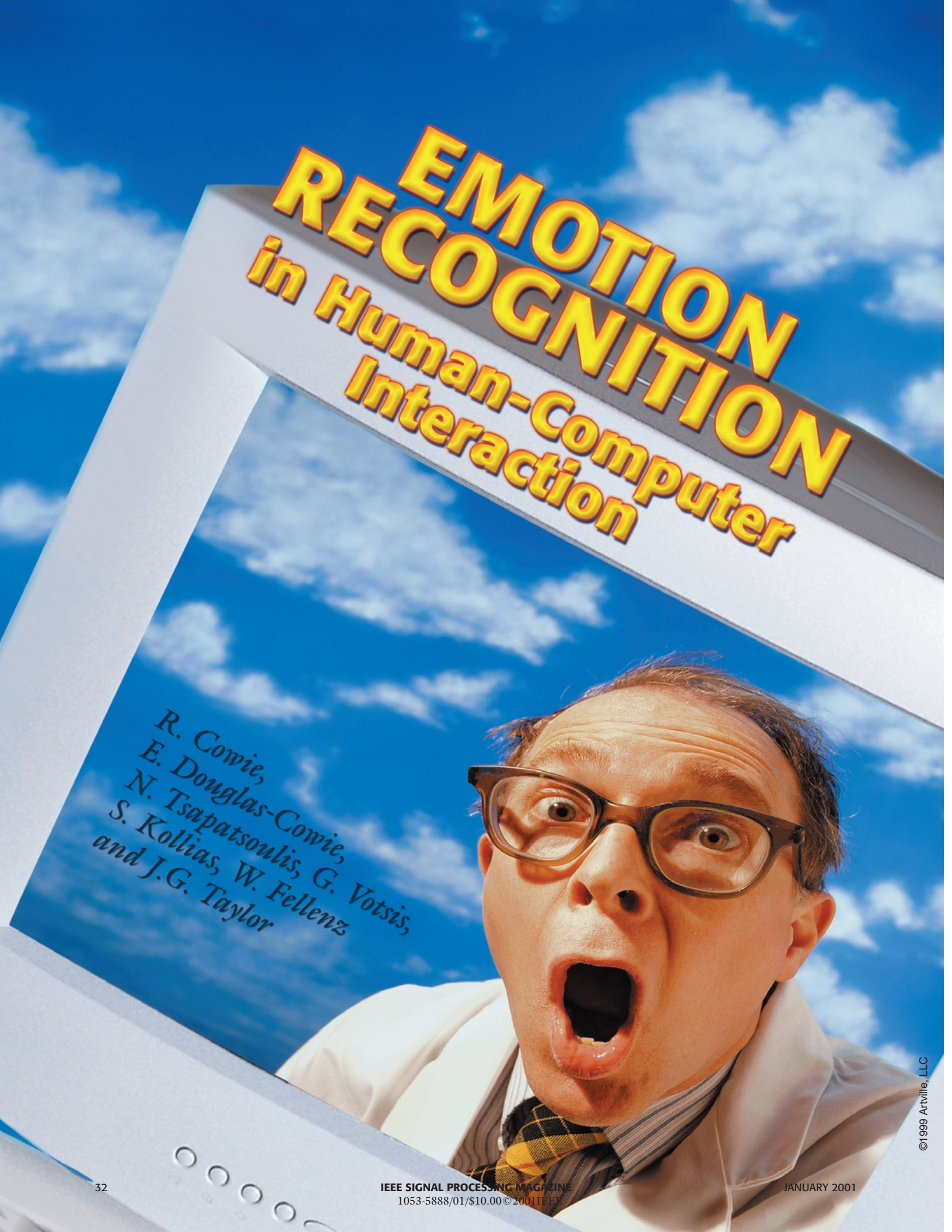
Emotion and Affective Computing

- 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. (**required**)
- Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S., "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, Jan. 2009. (**optional**)



1053-5888/01/$10.00©2001IEEE

Two channels have been distinguished in human interaction [1]: one transmits explicit messages, which may be about anything or nothing; the other transmits implicit messages about the speakers themselves. Both linguistics and technology have invested enormous efforts in understanding the first, explicit channel, but the second is not as well understood. Understanding the other party's emotions is one of the key tasks associated with the second, implicit channel. To tackle that task, signal processing and analysis techniques have to be developed, while, at the same time, consolidating psychological and linguistic analyses of emotion. This article examines basic issues in those areas. It is motivated by the PHYSTA project, in which we aim to develop a hybrid system capable of using information from faces and voices to recognize people's emotions.

The human sciences contain a bank of literature on emotion which is large, but fragmented. The main sources which are relevant to our approach are in psychology and linguistics, with some input from biology. Translating abstract proposals into a working model system is a rational way of consolidating that knowledge base. That approach has several attractions, particularly when referring to hybrid systems, which include symbolic and subsymbolic techniques.

First, building an emotion detection system makes it possible to assess the extent to which theoretical proposals explain people's everyday competence at understanding emotion. So long as it is technically impossible to apply that kind of test, theories can only be assessed against their success or failure on selected examples, and that is not necessarily a constructive approach.

Second, model building enforces coherence. At a straightforward level, it provides a motivation to integrate information from sources that tend to be kept separate. It can also have subtler effects, such as showing that apparently meaningful ideas are actually difficult to integrate, that conjunctions which seem difficult are quite possible, or that verbal distinctions and debates actually reduce to very little.

Hybrid systems have a particular attraction in that they offer the prospect of linking two types of elements that are prominent in reactions to emotion—articulate verbal descriptions and explanations and responses that are felt rather than articulated, which it is natural to think of as subsymbolic.

Another related major issue is the emergence of meaning from subsymbolic operations. Intuitively, meanings related to emotion seem to straddle the boundary between the logical, discrete, linguistic representations that classical computing handles neatly (perhaps too neatly to model human cognition well), and the fuzzy, subsymbolic representations that, for example, artificial neural networks construct. That makes the domain of emotion a useful testbed for technologies which aim to create a seamless hybrid environment, in which it is possible for something that deserves the name meaning to emerge.

The Implicit Channel

The implicit channel is a major feature of human communication, and if progress is made towards reproducing it, then applications can be expected to follow. It is useful, however, to indicate the kinds of application that can easily be foreseen. In particular, applications that provide a context against which to assess the likely relevance of different theoretical approaches. Obvious possibilities can be summarized under nine headings, beginning with broad categories and then considering more specific applications.

Convergence

It is a feature of human communication that speakers who are in sympathy, or who want to indicate that they are, converge vocally on a range of parameters [2]. Conversely, not to converge conveys a distinct message—roughly, aloofness or indifference. That is the message that is likely to be conveyed by an electronic speaker which always uses a register of controlled neutrality irrespective of the register used by a person interacting with it, and it is liable to interfere with the conduct of business. To vary its own register so that it can converge appropriately, a machine needs some ability to detect the speaker's state.

Interaction between Channels

The two channels of human communication interact: the implicit channel tells people "how to take" what is transmitted through the explicit channel. That becomes particularly critical in the context of full-blown conversation rather than minimal, stereotyped exchanges. There is a growing body of knowledge on the way prosody contributes to that function [3], and it is reasonable to see it as part of a wider domain linked to speaker state. For example, the same words may be used as a joke, or as a genuine question seeking an answer, or as an aggressive challenge (e.g., "I suppose you think England is going to win the World Cup"). Knowing what is an appropriate continuation of the interaction depends on detecting the register that the speaker is using, and a machine communicator that is unable to tell the difference will have difficulty managing conversation.

Augmenting Human Judgment

Some of the most immediate applications involve gathering data about signs of emotion available to a human, who is engaged in making judgments about another person and who wants to make them more accurately or objectively. The classical example is lie detection. Improving on human performance in that area is a tall order. There are areas, however, where augmentation is a real possibility. Two examples can be easily discerned. First, some clinical diagnoses depend on detecting vocal signs of emotion, such as the diagnosis of flattened affect in schizophrenia, which is an indicator of poor prognosis

Hybrid systems have a particular attraction in that they link two types of elements that are prominent in reactions to emotion—articulate verbal descriptions and explanations and responses that are felt rather than articulated.

and potential hospitalization [4]. Relying on psychiatrists' unaided judgment in that area may not be optimal, since they are not necessarily chosen for the sensitivity of their ears. Hence it makes sense to supplement their subjective impressions with relevant objective measures, and there is prima facie evidence that the technology for obtaining relevant measures is within reach [5]. Second, providers of teleconferencing are interested in on-screen displays carrying information about participants' emotional states to losses of sensitivity that result from unnaturalness of the medium.

Deconfounding

The vocal signs of emotion occupy what has been called the augmented prosodic domain [7]—a collection of features involving pitch (which is usually equated with fundamental frequency, abbreviated as F0), amplitude, the distribution of energy across the spectrum, and some aspects of timing [8]. Difficulties arise because the same domain carries other types of information, such as information about the stage an interaction is at (preliminary exchanges, business, inviting closure, signing off) [9]. It is important to develop ways of using these pieces of information to negotiate human/computer transactions and that depends on understanding emotion-related variation well enough to recognize which underlies a particular pattern in the augmented prosodic domain.

Production

There is a good deal of interest in generating voices, which have appropriate emotional coloring [10], [11]. There is a duality between that problem and the problem of recognizing emotion in speech. In particular, techniques for learning the subtleties of emotional speech may provide a way of generating convincingly emotional speech. Similar points apply to the visual expression of emotion. The generation of synthetic agents characteristics, which attribute convincing expression, is crucial for virtual reality, natural-synthetic imaging, and human-computer interaction. A special case arises with compression techniques where there is the possibility of information about emotion being extracted, transmitted, and used to govern resynthesis. Note that resynthesis techniques, which fail to transmit information about emotion, have the potential to be catastrophically misleading.

Tutoring

An obvious application for emotion-sensitive machines is automatic tutoring. An effective tutor needs to know whether the user is finding examples boring or irritating or intimidating. As voice and camera inputs get widely used, it is realistic to ask how such tutors could be made sensitive to those issues.

Avoidance

A second obvious type of application involves machines acting as functionaries—personal assistants, information providers, receptionists, etc. There would be clear advantages if these machines could recognize when the human they were interacting with was in a state that they were not equipped to handle and then either close the interaction or hand it over to someone who was equipped to handle it.

Alerting

Related to the avoidance function is providing systems which can alert a user to signs of emotion that call for attention. Alerting may be necessary because the speaker and the person to be alerted are in different places (e.g., an office manager being alerted to problems with an interaction between one of several members of staff and a client, a ward nurse being alerted to a patient in distress) or because the speaker's attention is likely to be focused on other issues so that signs of emotion are overlooked (e.g., a physician talking to a patient who is not presenting their real concern, an academic adviser with a student who has undisclosed problems). The issue may be to alert people to their own emotions (for instance, so that signs of strain that might affect a critical operation or negotiation are picked up before damage is done).

Entertainment

Commercially, the first major application of emotion-related technology may well be in entertainment and game programs which respond to the user's state. There is probably an immense market for pets, friends, and dolls which respond even crudely to the owner's mood.

Many of those applications could be addressed in a piecemeal way. It seems likely, however, that genuinely satisfying solutions will depend on a solid theoretical base. The main concern of this article is with the development of that kind of base.

It is important for both theory and application to recognize that the term "emotion" has a broad and a narrow sense. The narrow sense refers to what might be called full-blown emotion, where emotion is (temporarily) the dominant feature of mental life—it preempts ordinary de-

liberation and directs people strongly towards a course of action driven by the emotion. The broad sense covers what might be called underlying emotion, which colors a person's thoughts and actions to a greater or lesser extent without necessarily seizing control. To avoid confusion, we use the phrase "emotional state" to describe any mental state where emotion—full blown or underlying—might reasonably be considered to play a central role. The term "affect" has a similar scope, but tends to be used in clinical contexts.

This article considers emotion in the broad sense. People sometimes object to that on the grounds that "emotion" strictly means full-blown emotion. That kind of argument is plainly irrelevant to deciding how information technology should regard underlying emotion. Its priorities are pragmatic, in the sense that it has to deal with emotion as it occurs in real settings. In that context, it would be difficult to justify a policy of ignoring underlying emotion. For instance, it is a serious limitation if an alerting or tutoring system is blind to signs of emotions like boredom or anger until they become full blown.

The structure of the article is as follows. In the next section, we introduce theoretical approaches to emotion and present the type of responses emotion recognition systems may provide. Then we discuss emotion-related signals in speech and the feature analysis techniques which are related to them. Following that, a similar examination of emotion-related signals in the face and of expression recognition techniques will be given. The final two sections discuss the availability of test material and summarize the state-of-the-art and the kinds of development that it is possible to envisage.

A Descriptive Framework for Emotional States

Constructing an automatic emotion recognizer depends on a sense of what emotion is. Most people have an informal understanding, but there is a formal research tradition which has probed the nature of emotion systematically. It has been shaped by major figures in several disciplines—philosophy (Rene Descartes), biology (Charles Darwin), and psychology (William James)—but it is convenient to call it the psychological tradition. Here we consider how ideas drawn from that tradition can be used.

In the context of automatic emotion recognition, understanding the nature of emotion is not an end in itself. It matters mainly because ideas about the nature of emotion shape the way emotional states are described. They imply that certain features and relationships are relevant to describing an emotional state, distinguishing it from others, and determining whether it qualifies as an emotional state at all. A descriptive framework embodies judgments about what those features and relationships are, and a satisfactory one allows information about them to be set out in a systematic, tractable way.

It is useful to highlight three kinds of issues that hinge directly on the choice of a descriptive framework. First, there are domain issues: what is to be considered an emotional state? Second, there are input issues: what kinds of evidence warrant conclusions about emotional states? Third, there are output issues: what kind of information is it appropriate for an automatic emotion recognizer to deliver?

The role of the psychological tradition should not be overstated. Its conclusions are rarely firm enough to dictate the conceptual framework of automatic emotion recognition. On the other hand, it has generated a wealth of ideas and techniques with the potential to be useful if they are used in a pragmatic way. This section aims to introduce those ideas. We briefly review the tradition as a whole and then consider elements that relate directly to artificial emotion recognition, using the division between input and output issues, and separating out those that deal with physiological underpinnings. Domain issues are reviewed at the end.

The Psychological Tradition: A Selective Overview

The psychological tradition has been shaped by a few enduring ideas [12]. Descartes introduced the idea that a few basic emotions underlie the whole of emotional life (the term primary reflects a specific theory about them). Darwin [76] introduced the idea that emotions are inseparable from serviceable associated habits, i.e., distinctive action patterns selected by evolution because of their survival value. His main examples were facial expressions and bodily movements that signal emotion in humans and (he argued) animals. James drew attention to the intimate connection between emotion and somatic arousal. Arnold [18] emphasized that emotion entailed a cognitive appraisal, which alerted the organism to situations with certain special kinds of significance—in her own phrase, a "direct, immediate sense judgment of weal or woe" [18, p. 171].

Textbooks tend to present a standard view, which is that those ideas are all valid, and that between them, they capture the main aspects of emotion. An important implication is that emotions are syndromes, defined by the co-occurrence of several types of events, and it would be a misrepresentation to regard one aspect as the real substance of emotion (e.g., arousal) and the others as secondary. Several recent accounts propose that emotions reconfigure the organism—multiple systems lock together into a pattern that has evolved to deal efficiently with a particular, pressing kind of situation (e.g., [16], [184]). Oatley and Johnson-Laird [13] add the idea that emotional reconfiguration constitutes an information processing strategy that is potentially valuable for artificial agents as well as humans (Star Trek viewers may recognize the idea).

The standard view has real strengths, but it also faces unresolved problems. Three broad areas of difficulty can be distinguished, and all of them are relevant to automatic

emotion recognizers indicating areas where available science doesn't dictate how to proceed and pragmatic judgments must be made.

First, there is no agreement on a set of basic emotions. Even the criteria for choosing one set rather than another are not agreed upon, and in fact, focusing on different aspects of emotion notoriously tends to produce different lists. Considering the effort that has been devoted to the issue, that lack of convergence suggests that there may well be no natural units to be discovered. The immediate implication is that selecting lists of emotions to be recognized requires pragmatic choices.

Second, there is unresolved tension between the standard emphasis on emotions as a product of evolution and evidence that they are culture dependent. For instance, syndromes may function as recognizable emotions in one culture but not others. Well-known examples are the medieval European concept of *accidie* [30] and the Japanese concept of *amae* [29], roughly translated as sinful lethargy and "sweet dependence." Social constructivists, such as Averill and Harre, infer that "emotions are transitory social roles—that is, institutionalised ways of interpreting and responding to particular classes of situations" [28, p. 100]. Again, the immediate implication is that pragmatic choices need to be made about the way to handle cultural differences.

Third, accounts that emphasize discrete syndromes apply best to full-blown emotion. As a result, underlying emotion tends to be rather marginalized. Some authors justify that emphasis by appeal to the "strict" meaning of the word emotion. That rests on a model of word meaning that has been strongly criticized [24]. Once again, approaches to underlying emotion are a matter of pragmatic choice.

The outline above is only intended to provide a context for the descriptions that follow. More complete summaries are available, usually from a particular viewpoint (e.g., [14], [12], [17], and [25]).

Input-Related Issues

There is a large literature on the signs that indicate emotion, both within the psychological tradition and beyond it. Later on we look at technical aspects of the literature, but there are general issues useful to raise early on. Most fundamental is a point that has been made strongly by Russell in the context of facial expression [6]. Ecological validity has not been an overriding priority in research on the expression of emotions. As a result, it cannot be assumed that material from the research literature will transfer easily to real-world applications. Several kinds of complications need to be recognized.

One key group of issues relates to the idea, stemming from Darwin, that signs of emotion are rooted in biology and are therefore universal. In that vein, Ekman and his colleagues [104] have argued that there are universally recognized facial expressions for basic emotions, and biological considerations have been used to predict signs of emotion in the voice. A universal vocabulary of emotional signs, rooted in biological principles, is attractive for information technology. It is clear, however, that not all emotion is reflected in universally recognized signs. Several factors complicate the situation.

Constructing an automatic emotion recognizer depends on a sense of what emotion is.

Display Rules: Ekman and Friesen stressed the universal underpinnings of facial expression, but also the way culturally defined display rules are used to "manage the appearance of particular emotions in particular situations" [104]. Unrestrained expressions of anger or grief are strongly discouraged in most cultures and may be replaced by an attempted smile rather than a neutral expression; detecting those emotions depends on recognizing signs other than the universally recognized archetypal expressions.

Deception: There is a fine line between display rules and deception. Deliberately misrepresenting emotional states is manifestly part of social life, and for that reason, detecting deception has been a key application for the psychological tradition. It is another pragmatic decision on how artificial emotion recognizers should approach deception. Trying to detect it puts a premium on using indicators that humans do not (e.g., physiological). Our main interest is trying to match the things that humans can do, and that means accepting that the system will be deceived as humans are.

Systematic Ambiguity: Signs which are relevant to emotion may also have alternative meanings. Obviously, lowered eyebrows may signify concentration as well as anger. Less obviously, there are strong similarities between the prosodic characteristics associated with depression [189] and those associated with poor reading [190]. The systematic ambiguity of individual signs is a serious issue for any practical application. It makes coordinating information from different modalities a high priority; this is why later sections consider both speech and facial expression.

A second group of issues relates to the structure of information sources. A large proportion of research has dealt with sources that can be thought of as qualitative targets—a smile, a distinctive kind of pitch contour in speech, and so on. There are practical reasons for that emphasis, particularly in research without access to high technology: qualitative targets are comparatively easy to identify or manipulate.

It has gradually become clear that structurally different types of sources need to be considered. Some sources function as gestures, which are extended in time: for instance, judgments about a smile depend on its time course as well as its final shape. Other cues appear to lie in the manner of an action, for instance, the way spoken words are stressed [8]. Others are heavily dependent on context for their meaning—for instance, a flush that in isolation

Table 1. Emotion Words from Whissell and Plutchik.							
	Activ	**Eval**	**Angle**		**Activ**	**Eval**	**Angle**
Accepting			0	Disgusted	5	3.2	161.3
Adventurous	4.2	5.9	270.7	Disinterested	2.1	2.4	127.3
Affectionate	4.7	5.4	52.3	Disobedient			242.7
Afraid	4.9	3.4	70.3	Displeased			181.5
Aggressive	5.9	2.9	232	Dissatisfied	4.6	2.7	183
Agreeable	4.3	5.2	5	Distrustful	3.8	2.8	185
Amazed	5.9	5.5	152	Eager	5	5.1	311
Ambivalent	3.2	4.2	144.7	Ecstatic	5.2	5.5	286
Amused	4.9	5	321	Elated			311
Angry	4.2	2.7	212	Embarrassed	4.4	3.1	75.3
Annoyed	4.4	2.5	200.6	Empty	3.1	3.8	120.3
Antagonistic	5.3	2.5	220	Enthusiastic	5.1	4.8	313.7
Anticipatory	3.9	4.7	257	Envious	5.3	2	160.3
Anxious	6	2.3	78.3	Exasperated			239.7
Apathetic	3	4.3	90	Expectant			257.3
Apprehensive			83.3	Forlorn			85
Ashamed	3.2	2.3	83.3	Furious	5.6	3.7	221.3
Astonished	5.9	4.7	148	Generous			328
Attentive	5.3	4.3	322.4	Gleeful	5.3	4.8	307
Awed			156.7	Gloomy	2.4	3.2	132.7
Bashful	2	2.7	74.7	Greedy	4.9	3.4	249
Bewildered	3.1	2.3	140.3	Grief-stricken			127.3
Bitter	6.6	4	186	Grouchy	4.4	2.9	230
Boastful	3.7	3	257.3	Guilty	4	1.1	102.3
Bored	2.7	3.2	136	Happy	5.3	5.3	323.7
Calm	2.5	5.5	37	Helpless	3.5	2.8	80
Cautious	3.3	4.9	77.7	Hesitant			134
Cheerful	5.2	5	25.7	Hopeful	4.7	5.2	298

(Continued on next page)

Table 1. Emotion Words from Whissell and Plutchik *(continued)*.							
	Activ	**Eval**	**Angle**		**Activ**	**Eval**	**Angle**
Confused	4.8	3	141.3	Hopeless	4	3.1	124.7
Contemptuous	3.8	2.4	192	Hostile	4	1.7	222
Content	4.8	5.5	338.3	Humiliated			84
Contrary	2.9	3.7	184.3	Impatient	3.4	3.2	230.3
Co-operative	3.1	5.1	340.7	Impulsive	3.1	4.8	255
Critical	4.9	2.8	193.7	Indecisive	3.4	2.7	134
Curious	5.2	4.2	261	Indignant			175
Daring	5.3	4.4	260.1	Inquisitive			267.7
Defiant	4.4	2.8	230.7	Interested			315.7
Delighted	4.2	6.4	318.6	Intolerant	3.1	2.7	185
Demanding	5.3	4	244	Irritated	5.5	3.3	202.3
Depressed	4.2	3.1	125.3	Jealous	6.1	3.4	184.7
Despairing	4.1	2	133	Joyful	5.4	6.1	323.4
Disagreeable	5	3.7	176.4	Loathful	3.5	2.9	193
Disappointed	5.2	2.4	136.7	Lonely	3.9	3.3	88.3
Discouraged	4.2	2.9	138	Meek	3	4.3	91
Nervous	5.9	3.1	86	Self-conscious			83.3
Obedient	3.1	4.7	57.7	Self-controlled	4.4	5.5	326.3
Obliging	2.7	3	43.3	Serene	4.3	4.4	12.3
Outraged	4.3	3.2	225.3	Shy			72
Panicky	5.4	3.6	67.7	Sociable	4.8	5.3	296.7
Patient	3.3	3.8	39.7	Sorrowful	4.5	3.1	112.7
Pensive	3.2	5	76.7	Stubborn	4.9	3.1	190.4
Perplexed			142.3	Submissive	3.4	3.1	73
Planful			269.7	Surprised	6.5	5.2	146.7
Pleased	5.3	5.1	328	Suspicious	4.4	3	182.7
Possessive	4.7	2.8	247.7	Sympathetic	3.6	3.2	331.3
Proud	4.7	5.3	262	Terrified	6.3	3.4	75.7

(Continued on next page)

Table 1. Emotion Words from Whissell and Plutchik *(continued)*.							
	Active	**Eval**	**Angle**		**Active**	**Eval**	**Angle**
Puzzled	2.6	3.8	138	Timid			65
Quarrelsome	4.6	2.6	229.7	Tolerant			350.7
Ready			329.3	Trusting	3.4	5.2	345.3
Receptive			32.3	Unaffectionate	3.6	2.1	227.3
Reckless			261	Uncertain			139.3
Rebellious	5.2	4	237	Uncooperative			191.7
Rejected	5	2.9	136	Unfriendly	4.3	1.6	188
Remorseful	3.1	2.2	123.3	Unhappy			129
Resentful	5.1	3	176.7	Unreceptive			170
Revolted			181.3	Unsympathetic			165.6
Sad	3.8	2.4	108.5	Vascillating			137.3
Sarcastic	4.8	2.7	235.3	Vengeful			186
Satisfied	4.1	4.9	326.7	Watchful			133.3
Scared			66.7	Wondering	3.3	5.2	249.7
Scornful	5.4	4.9	227	Worried	3.9	2.9	126

might signal either pleasure or anger. It is not clear how much the everyday expression of emotion relies on traditional qualitative targets and how much on these subtler types of source. That theme runs through later sections.

Output-Related Issues

The obvious goal for an automatic recognizer is to assign category labels that identify emotional states. However, labels as such are very poor descriptions. In addition, humans use a daunting number of labels to describe emotion. Table 1 illustrates the point using two lists cited by Plutchik, who takes a relatively conservative view of what emotion is. It is difficult to imagine artificial systems beginning to match the level of discrimination that those lists imply. This section considers alternative ways of representing emotional states, drawn from the psychological tradition. It begins by describing representations that are simple and uniform, but approximate, and moves on to more complex options.

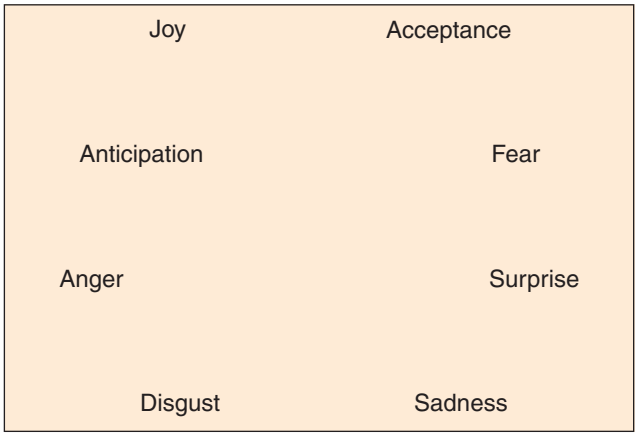
Activation-Evaluation Space and Related Representations

Activation-emotion space is a representation that is both simple and capable of capturing a wide range of significant issues in emotion. It rests on a simplified treatment of two key themes.

Valence: The clearest common element of emotional states is that the person is materially influenced by feelings that are valenced, i.e., they are centrally concerned with positive or negative evaluations of people or things or events. The link between emotion and valencing is widely agreed, although authors describe it in different terms. Arnold refers to the "judgment of weal or woe" [18, p. 171]; Tomkins, describes affect as what gives things value—"without its amplification, nothing else matters, and with its amplification, anything else can matter" [26, pp. 355-356]; Rolls sees emotional processing as where "reward or punishment value is made explicit in the representation" [182, p. 6].

Activation Level: Research from Darwin forward has recognized that emotional states involve dispositions to act in certain ways. A basic way of reflecting that theme turns out to be surprisingly useful. States are simply rated in terms of the associated activation level, i.e., the strength of the person's disposition to take some action rather than none.

The axes of activation-evaluation space reflect those themes. The vertical axis shows activation level and the horizontal axis evaluation. A basic attraction of that arrangement is that it provides a way of describing emotional states which is more tractable than using words, but which can be translated into and out of verbal descrip-



▲ *1. Plutchik's "emotion wheel."*

tions. Translation is possible because emotion-related words can be understood, at least to a first approximation, as referring to positions in activation-emotion space. Various techniques lead to that conclusion, including factor analysis, direct scaling, and others [22].

The numbers in Table 1 show how words can be related to activation-evaluation space. It reflects two studies, one due to Whissell [22] and the other due to Plutchik [23]. The first two numerical columns show values for activation and evaluation that were found in the study by Whissell. Terms like patient and cautious (at 3.3) signify a mid level of activation, surprised and terrified (over 6) signify a high level, and bashful and disinterested (around 2) signify low activation. Evaluation ranges from guilty (at 1.1), representing the negative extreme, to delighted (at 6.6), representing the positive extreme.

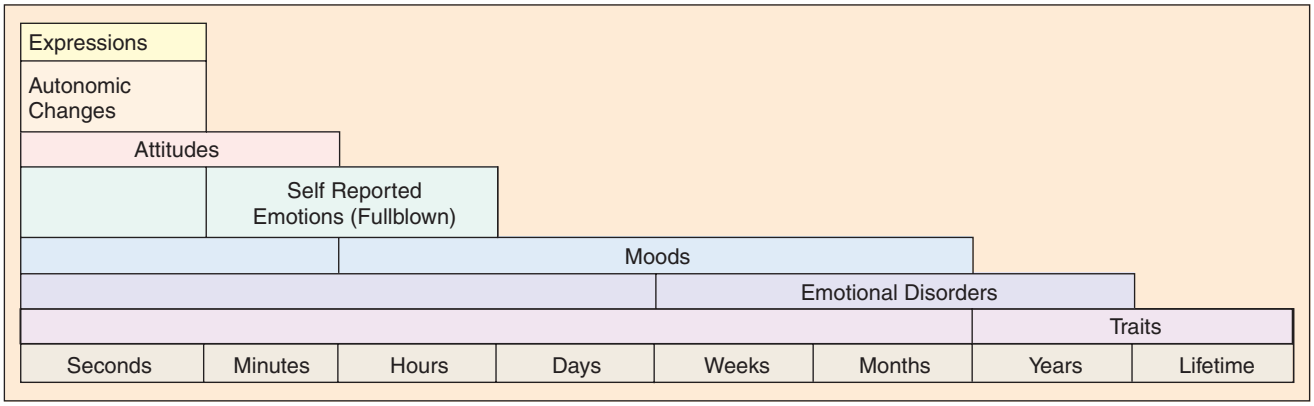
A surprising amount of emotional discourse can be captured in terms of activation-emotion space. The third column of Table 1, based on data from Plutchik, reflects one development. Words that describe full-blown emotions are not evenly distributed in activation-emotion space. Instead they tend to form a roughly circular pattern. From that and related evidence, Plutchik has argued that there is a circular structure inherent in emotionality. The figures in the third column of Table 1 reflect that approach: they give empirically derived angular measures which specify where on the emotion circle each word lies, in terms of a reference system defined by axes running approximately from acceptance (0) to disgust (180) and from apathetic (90) to curious (270). The measures can be called emotional orientation.

Identifying the center as a natural origin has several implications. Emotional strength can be measured as the distance from the origin to a given point in activation-evaluation space. The concept of a full-blown emotion can then be translated roughly as a state where emotional strength has passed a certain limit. An interesting implication is that strong emotions are more sharply distinct from each other than weaker emotions with the same emotional orientation. A related extension is to think of primary or basic emotions as cardinal points on the periphery of an emotion circle. Plutchik has offered a useful formulation of that idea, the "emotion wheel" which is shown in Fig. 1.

Activation-evaluation space is a surprisingly powerful device, and it has been increasingly used in computationally oriented research (e.g., [96], [183], [197]). It has to be emphasized, however, that representations of that kind depend on collapsing the structured, high-dimensional space of possible emotional states into a homogeneous space of two dimensions. There is inevitably loss of information and, worse still, different ways of making the collapse lead to substantially different results. That is well illustrated in the fact that fear and anger are at opposite extremes in Plutchik's emotion wheel, but close together in Whissell's activation/emotion space. Extreme care, thus, is needed to ensure that collapsed representations are used consistently. That point is followed up later on.

Categories Related to Time

Everyday usage divides emotional states into categories that are related to time, reflecting the fact that emotional life has a definite temporal structure. Emotion in its narrow sense—full-blown emotion—is generally short lived and intense. "Mood" describes an emotional state that is underlying and relatively protracted. Emotional traits are more or less permanent dispositions to enter certain emotional states. Emotional disorders—such as depression or pathological anxiety—also fall within the broad category of emotional states. They may involve both full-blown and protracted emotion. Figure 2, adapted from [17],



▲ *2. Temporal characteristics of emotion categories.*

summarizes the temporal characteristics of these categories. It is interesting that emotion words can refer to many categories, e.g., "happy" may describe a full-blown emotion, a mood, or a trait.

Dealing with the temporal structure of emotional life is a difficult challenge, but an interesting one. There are obvious reasons for trying to assess whether a user is briefly angry or constitutionally bad tempered. It is natural for an emotion recognizer to build up a representation of the user on at least two time scales, one short (dealing with full-blown emotions) and one long (dealing with moods and traits).

Table 2. Extract from "Emotions and Their Derivatives" (Plutchik, 1980).

Stimulus	Cognition	Subjective Language	Behavior
Threat	Danger	Fear	Escape
Obstacle	Enemy	Anger	Attack
Potential mate	Possess	Joy	Mate
Loss of valued individual	Abandonment	Sadness	Cry
Member of one's group	Friend	Acceptance	Groom
Unpalatable object	Poison	Disgust	Vomit
New territory	What's out there?	Expectation	Map
Unexpected object	What is it?	Surprise	Stop

Appraisals—Expanding the Evaluation Axis

Cognitive theories [19] stress that emotions are intimately connected to a situation that is being experienced or imagined by the agent—or, more precisely, to mental representations that highlight key elements of a situation and identify them as positive or negative. These representations have generally been called appraisals. An appraisal can be thought of as a model which is selective and valenced—i.e., highlights key elements of a situation and their values for good or ill.

Evaluation level amounts to an extremely reduced description of the appraisal that the agent has in mind. Richer analysis of an emotion depends on expanding that description by specifying what is appraised as good or bad about the situation. The psychological tradition offers various ways of making that expansion. For information technology, those expansions offer ways of associating an emotional state with a meaningful description rather than an unanalyzed label.

One broad approach follows up Darwin's view of emotions as evolutionary adaptations. Table 2, due to Plutchik, illustrates that approach. The first two columns effectively identify basic appraisals that could apply to humans or animals alike. The problem with that kind of analysis is that even if it does capture something about the roots of emotion, it is not obvious how it could be transferred to any situation where an automatic emotion recognizer would be likely to operate. That is linked to the fact that the scheme is oriented towards full-blown emotions, and extrapolation to underlying emotion is nontrivial.

In contrast, cognitive approaches have tried to set out underlying distinctions from which it is possible to derive a range of appraisals corresponding to the range of emotions. Table 3 summarizes the distinctions used in two substantial proposals of that kind, due to Roseman [193], [194] and to Ortony et al. [195]. Both include two distinctions that can be regarded as basic—whether the key elements of the situation are positively or negatively evaluated in themselves and whether they make the agent's goals more or less likely to be achieved. Roseman identified additional distinctions based on the way agents appraise key elements of the perceived situation—what agency is responsible for it, whether the main elements are known or unknown, and whether the agent regards him- or herself as powerful or powerless. Ortony et al. advocated a different kind of distinction, based on the idea

Table 3. Distinctions from which Appraisals Corresponding to a Range of Emotions Can Be Derived.

	Roseman	Ortony et al.
What are the elements that form the focus of the appraisal?		
Focal agent (self other(s))		+
Focal level (objects/actions/consequences of actions or events)		+
How are focal elements evaluated?		
Intrinsic value (intrinsically appealing/aversive)	+	+
Contextual value (consistent/inconsistent with agent's aspirations)	+	+
Clarity (known/uncertain/unknown)	+	
Agency (caused by self/other agent/circumstances)	+	

that appraisals may emphasize different kinds of element. The focus may be on different agents—the person experiencing the emotion or someone else. It may also be on different levels—objects (which may be people or things), actions (of people or animals), or sequences of causally related actions or events. Broadly speaking, the range of emotions that can be associated with an object as such is much narrower than the range of emotions that can be related to a sequence of events involving oneself and various others.

Cognitive theory provides a strong argument against equating emotion recognition with assigning category labels. Instead, it involves modeling the way a person perceives the world (or key aspects of it). Labeling could still be pivotal if the process proceeded by assigning an unanalyzed category label and accessing the relevant appraisal via the label. It clearly makes sense, however, to consider the opposite route. The key question is whether there could be signs that a person perceives him/herself as strong or weak, that he/she feels in possession of adequate information, or that he/she is considering a sequence of events extending into the past and/or the future rather than being focused simply on the present. If so, then identifying those signs could precede and facilitate assigning category labels.

It seems likely to be a long time before automatic emotion recognizers will incorporate all the subtleties considered by Ortony et al. Analyses like theirs, however, help to highlight issues that a simpler approach might address. They also identify terms that may be difficult to attribute automatically; some appraisals may involve multiple agents and complex relationships, such as pity, remorse, or gratitude.

Table 4. Basic Action Tendencies Identified by Frijda.

Action Tendency	Function	Emotion
Approach	Permits consummatory behavior	Desire
Avoidance	Protection	Fear
Being-with	Permits consummatory activity	Enjoyment, confidence
Attending	Orientation to stimuli	Interest
Rejecting	Protection	Disgust
Nonattending	Selection	Indifference
Agonistic (attack/threat)	Regaining control	Anger
Interrupting	Reorientation	Shock, surprise
Dominating	Generalized control	Arrogance
Submitting	Secondary control	Humility, resignation

Action Tendencies—Expanding the Activation Axis

The activation axis of activation-evaluation space can be expanded by distinguishing the kinds of action that emotion may prompt or inhibit. Frijda, in particular, has explored the link between emotions and activation tendencies. Table 4 describes a set of basic action tendencies identified by Frijda [196]. He has used that kind of correspondence to argue that emotions and their associated action tendencies are "one and the same thing." The claim is plausible for the examples in the table, but harder to sustain elsewhere. Joy, he himself calls an activation mode; it means generalized readiness to take opportunities for activity. The link to action is still less obvious in cases such as pity or remorse.

Clearly developing good descriptions of action tendencies is an important goal for information technology: after all, the usefulness of automatic emotion recognizers depends on their ability to anticipate what people may do when they are in particular emotional states. It is less clear how the descriptions might be generated. In the style of description developed by Frijda, action tendencies are more or less discrete, and there seems little option but to label the emotional state and access the relevant action tendency via the label. It would be interesting if possible actions could be decomposed in a way that allowed recognizable signs to be associated with broad types of action tendency and used to identify the emotion.

An example of that kind of decomposition comes from research on the development of emotions. Figure 3 shows a scheme put forward by Fox [27], which has a strong flavor of action tendencies. It proposes that emotions originate in two broad action tendencies—to approach or to withdraw. These are differentiated at the second level into approach with a view to gaining pleasure (joy), approach with a view to gaining information (interest), and approach with a view to confrontation (anger)—and so on.

Another variation on the theme of action tendencies is to consider emotions as prompts to plan an action pattern. For Oatley and Johnson-Laird [13], [20], the fact that emotions tend to be signaled reflects the fact that establishing a plan of action may not be an individual matter, because the viability of a given plan may depend on other people's attitudes. The idea is attractive, but like sophisticated appraisal models, its relevance to information technology may be in the long term.

Category Labels

Category labels are not a sufficient representation of emotional state, but they are probably necessary. The choice of a suitable set to use is important, but not straightforward.

The sheer variety of emotion terms that exist in everyday language was illustrated in Table 1. There is no immediate prospect of an automatic emotion recognizer using the whole range. Worse, the diversity of terms is an obstacle to synthesis. As will be shown in the next section, research in some traditions has made a point of finding words that convey the exact shade of emotion that is experienced or conveyed in particular situations. The result is a mass of material that defies integration because no two studies talk about the same emotional category. Converging on a smaller vocabulary is a prerequisite for progress.

		1st Level			
	Approach			Withdrawal	
		2nd Level			
Joy	Interest	Anger	Distress	Disgust	Fear
		3rd Level			
Pride	Concern	Hostility	Misery	Contempt	Horror
Bliss	Responsibility	Jealousy	Agony	Resentment	Anxiety

▲ *3. Development of emotional distinctions (after Fox).*

The traditional approach has been to expect that science will define a core vocabulary corresponding to basic emotions. At this stage, that seems optimistic. Focusing on different aspects of emotion suggests different sets of basic emotions or none; see, e.g., the schemes reported above from Plutchik (Fig. 1), Frijda (Table 4), Fox (Fig. 3), and the Ortony et al. proposal which implies an indefinite number of possible appraisal types. There is still less encouragement for the Cartesian idea that basic categories can serve as primaries, mixing to produce other emotions—that has been called the "palette theory."

Some categories do appear on almost every list of basic emotions—happiness, sadness, fear, anger, surprise, and disgust. It is not in doubt that they are key points of reference. It is probably best to describe them as archetypal emotions, which reflects the fact that they are undeniably the obvious examples of emotion. Although the archetypal emotions are important, they cover rather a small part of emotional life. It is a pragmatic problem to find a set of terms that covers a wider range without becoming unmanageable. We have developed a pragmatic approach to finding such a set, which we call a basic emotion vocabulary. It is reported below.

BEEVer—A Pragmatic Synthesis

Recently, ideas from the psychological tradition were used to assemble a descriptive framework suitable for use in automatic emotion recognition [197]. Category terms were chosen by asking naive subjects to select, from a longer list, 16 words that would form a basic English emotion vocabulary (BEEV for short). The meaning that they attached to each word was assessed in two ways: 1) subjects located it in activation-emotion space and 2) answered questions about the broad kind of action tendency and appraisal that the word implied.

The questions about action tendency followed Fox: subjects rated their expectation that an individual in that state would tend to engage with the person or situation causing the emotion, withdraw, or try to acquire information. Following Ortony et al. [195], the appraisal questions tried to tap implied sequences of events. They asked whether the individual would be concerned with a situation in his or her current surroundings and/or with one in the past, one in the future, or one that was primarily in his or her own mind. Each situation that was deemed relevant was then considered, asking about the individual's perception of his/her own power, knowledge or lack of it, and moral justification, along with parallel questions about any significant other in the situation.

The main results are summarized in Table 5. The words in the first column were selected from an initial list (based on Table 1 and similar sources) by removing those that were almost never included in a BEEV. The words in bold type were included by at least half the subjects. The next two columns show locations in activation-emotion space, giving first emotional orientation and then emotion strength (1=maximum). The words are ordered by emotional orientation. The remaining columns highlight features that provided additional discrimination. The "approach- withdraw" responses showed a default pattern: negative emotional orientation generally meant a balance in favor of withdrawal, and positive meant a balance in favor of engaging. The interesting cases are the exceptions. Worry, anger, sympathy, and surprise showed no decisive balance, i.e., action was judged unpredictable. Boredom, disgust, fear, and anxiety were marked by an unusually strong inclination to withdraw (shown by the + prefix). The tendency to seek information was not usually an issue. Cases where it was are coded in terms of an open mind (disposed to seek information) or a closed one. The remaining columns show the individual's own perceived power in the present situation, and his/her perceived orientation to situations in the present, past and future, and elsewhere (usually imagined).

The data are presented not as a reference (a larger subject pool would be needed for that), but to illustrate the kind of descriptive system that it is natural to develop. It makes sense to understand that assigning a category term such as "worried" means, roughly, that the user has a generally negative outlook, feels powerless and in need of information, is concerned with the future, and might act unpredictably. That kind of information has the potential both to guide response and to suggest an explanation for signals indicating qualities such as powerlessness and need of information.

Physiological Issues

Although physiology has been an integral part of the psychological tradition, it interacts relatively little with the

Table 5. A Basic English Emotion Vocabulary with Simple Semantic Features.								
	Emotional Orientation	Strength of Emotion	Disposed to Engage or Withdraw	Open- or Closed-minded	Own Perceived Power	Oriented to Surroundings	Oriented to Other Time	Oriented Elsewhere
Bored	–166	.46	+withdraw	++closed	–power	+surround	–past –future	
Disappointed	–133	.49	withdraw				+past	
Guilty	–126	.53	withdraw				+past	
Desparing	–116	.99	withdraw		–power			
Hurt	–115	.75	withdraw					
Sad	–101	.78	withdraw		–power	+surround		
Ashamed	–95	.74	withdraw				+past	
Resentful	–89	.74	withdraw				+past	
Jealous	–89	.48	withdraw			+surround		
Worried	–80	.65	unpredict-able	+open	–power		+future	
Disgusted	–76	.73	+withdraw	++closed		+surround		
Disagreeable	–67	.47	withdraw					–elsewh
Annoyed	–62	.5	withdraw			+surround		–elsewh
Irritated	–61	.64	withdraw	+closed		+surround		
Disapproving	–57	.31	withdraw					–elsewh
Embarrassed	–55	.42	withdraw	+closed				
Afraid	–52	.84	+withdraw	+closed	– –power		+future	
Angry	–48	.95	+unpredict-able			+surround		
Anxious	–43	.72	+withdraw		–power		+future	+elsewh
Nervous	–38	.68	withdraw			+surround	+future	+elsewh
Panicky	–25	.86	withdraw	+closed	–power		–past	
Sympathetic	5	.66	unpredict-able	+open				
Surprised	7	.78	+unpredict-able				–past	
Interested	17	.7	engage	+open				
Excited	24	.95	engage	+open			+future	

(Continued on next page)

Table 5. A Basic English Emotion Vocabulary with Simple Semantic Features *(continued).*								
	Emotional Orientation	Strength of Emotion	Disposed to Engage or Withdraw	Open- or Closed-minded	Own Perceived Power	Oriented to Surroundings	Oriented to Other Time	Oriented Elsewhere
Loving	25	.84	+engage					
Affectionate	43	.72	engage					
Pleased	44	.52	engage					
Confident	44	.75	engage		+power			
Happy	47	.71	engage					
Joyful	52	.66	+engage			+surround	–future	–elsewh
Amused	53	.71	engage			+surround	–past, –future	
Proud	66	.59	engage		+power		–future	
Hopeful	69	.74	engage	+open			++future	+elsewh
Calm	115	.72	engage				–past	+elsewh
Content	136	.66	engage					+elsewh
Relieved	149	.75	unpredict-able				+past	+elsewh
Serene	151	.9	engage			+surround		+elsewh
Relaxed	153	.68	engage	+open			–past	+elsewh
Satisfied	172	.62	engage			+surround	–past	+elsewh

main issues considered in this article. For completeness, key ideas are summarized briefly here.

Emotional arousal has a range of somatic correlates, including heart rate, skin resistivity, temperature, pupillary diameter, and muscle activity. These have been widely used to identify emotion-related states—most obviously in lie detection. Traditional versions of that approach have an obvious limitation, in that they require the user to be wired. In addition, the most extensively studied application, lie detection, gives unreliable results, even in the hands of a skilled human operator [198]. An ingenious variant with applications in human computer interaction (HCI) is to measure some parameters using a specially constructed mouse (http://www.almaden.ibm.com/cs/blueeyes/ mouse.html#top). Speech has also been treated as a source of evidence for somatic changes associated with emotion.

Research on brain mechanisms of emotion is a rapidly growing field (for a recent review, see [182]). Its main relevance here is that it reinforces the view that emotion is multifaceted. A number of different brain systems are strongly associated with emotion, and they map at least roughly onto the kind of division that has been outlined above. The amygdala have rich inputs from sensory systems and are involved in learning the reward values of stimuli. It is natural to interpret them as a key site in valenced appraisal. The orbitofrontal cortex is involved in preparing behavioral responses and also in some autonomic responses. It is natural to link its function to action tendencies. The basal forebrain has widespread effects on cortical activation, and direct links to autonomic nuclei; the fact suggests a role in arousal. Various cortical areas seem to have emotion-related functions. Neurons in the superior temporal sulcus respond to emotionally significant facial expressions. There also appear to be differences between the two cerebral hemispheres, relating perhaps to visual recognition as against verbal articulation of emotion.

It is striking how extensive emotion-related systems are in the brain. That underlines the importance of emotion in human life and also the scale of the task that is likely to be involved in developing an artificial emotion recognizer which matches human abilities.

Overview and Domain Issues Revisited

The domain of emotional states is considered at the end of this section because ideas about its structure and boundaries depend on ideas about representing emotion, which have now been outlined.

The representations that have been considered reflect the standard view that several types of element co-occur in full-blown emotions—appraisals, feelings, arousal, and action tendencies, including tendencies to give signs of emotion. If so, full-blown emotions would appear to belong within a wider domain of states that include some of those elements, but not necessarily all. That is a useful model for what we have called the domain of emotional states. We regard that as the natural topic for automatic emotion recognition. Among other things, that approach enshrines the reasonable principle that full-blown emotion ought to be considered in a context of states that might be confused with it, or lead to it, or have some of the same implications for behavior.

Several types of state that belong in the wider domain have been mentioned—moods, weak emotions, and states that may lead to full-blown emotions. Two others are worth highlighting.

Linguists, in particular, have studied states such as hostility, sarcasm, and curiosity, which are described as attitudes. There are signs that express these states, and they are distinctly valenced—i.e., positive or negative feelings towards something are a prominent element— but they are not strongly associated with arousal or concrete action tendencies. They tend to involve rather sophisticated kinds of appraisal. Some psychologists imply that attitude is a kind of affect [184], while others consider affect as a component of attitude [185]. The difference is not important here.

States such as distress, euphoria, or eagerness are partly similar. They are valenced and associated with rather nonspecific action tendencies. The appraisals are quite nonspecific, however, and they are strongly associated with arousal. Valenced arousal seems a suitable term for them. Both science and pragmatics suggest that states like those fall within the natural remit of automatic emotion recognition—science, because it would be unsatisfying to study full-blown emotions without reference to other states where the same elements play a key role (particularly the same kinds of sign), and pragmatics, because the states form a large part of the way emotional issues enter into everyday life.

Concentrating on archetypal emotions might be justified as a way into the wider domain. It is clear, though, that there are pitfalls on that route. Pure emotion is difficult to study because it is relatively rare and short lived, and eliciting it presents ethical problems. That makes it easy to slip into using simulations as a surrogate, and their ecological validity is highly suspect. There are also warning signs that it may be difficult to generalize findings from that approach. The reason may be that in emotion, as elsewhere, using carefully selected evidence makes it possible to evade issues that rapidly become important in anything but idealized cases. An aspect of that problem is that categorical representations may apply well to archetypal emotions, but much less so elsewhere.

Against that background, it makes sense to consider alternative strategies. Three in particular arise out of the representations that have been considered. Activation-evaluation space points to one alternative, which is to begin by assigning descriptions that are coarse, but that can be applied over a wide range of states. BEEVer embodies a related idea, which is to identify a reduced vocabulary that allows a wide range of states to be described at least roughly. The third alternative is to explore correspondences between expressive signs and features that run through a range of emotional states—for instance, the features in Table 5 that emerged in the BEEVer study. The reviews of sources that follow are guided by that assessment, taking a broad view of the subject matter, and of the ways we may approach it.

Table 6. Emotions and Speech Parameters (from Murray and Arnott, 1993).

	Anger	**Happiness**	**Sadness**	**Fear**	**Disgust**
Rate	Slightly faster	Faster or slower	Slightly slower	Much faster	Very much faster
Pitch Average	Very much higher	Much higher	Slightly lower	Very much higher	Very much lower
Pitch Range	Much wider	Much wider	Slightly narrower	Much wider	Slightly wider
Intensity	Higher	Higher	Lower	Normal	Lower
Voice Quality	Breathy, chest	Breathy, blaring tone	Resonant	Irregular voicing	Grumble chest tone
Pitch Changes	Abrupt on stressed	Smooth, upward inflections	Downward inflections	Normal	Wide, downward terminal inflects
Articulation	Tense	Normal	Slurring	Precise	Normal

Table 7. Table of Speech and Emotion (a).				
		Anger	**Happiness/Elation**	**Sadness**
Acoustic	Pitch	Increase in mean [156], [157], [158], median [32], range [32], [158], [159], variability [32], [160]	Increase in mean [156], [157], [161], [168], range [160], [166], [169], variability [160], [169]	Below normal mean F0 [158], [169], [171], range F0 [157], [158]
	Intensity	Raised [156], [158], [161], [162]	Increased [161], [170]	Decreased [156], [163]
	Duration	High rate [156], [163], [164], reduced rate [165]	Increased rate [156], [171], slow tempo [161]	Slightly slow [164], [173], long pitch falls [156]
	Spectral	High midpoint for av spectrum for nonfric portions [5]	Increase in high-frequency energy [161], [172]	Decrease in high-frequency energy [172], [174]
Contour		Angular frequency curve [157], Stressed syllables ascend frequently and rhythmically [166], irregular up and down inflection [156], level average pitch except for jumps of about a musical fourth or fifth on stressed syllables [166]	Descending line [166], melody ascending frequently and at irregular intervals [166]	Downward inflections [156]
Tone Based		Falling tones [37]		
Voice Quality		Tense [164], [166], breathy [167], heavy chest tone [167], blaring [156]	Tense [15], breathy [166], blaring [156]	Lax [15], resonant [156]
Other		Clipped speech [156], irregular rhythm basic opening and closing, articulatory gestures for vowel / consonant alternation more extreme [165]	Irregular stress distribution [166], capriciously alternating level of stressed syllables [166]	Slurring [156], rhythm with irregular pauses [156]

Speech and Emotional States

Speech consists of words spoken in a particular way. This section is concerned primarily with the information about emotion that resides in the way the words are spoken.

There is a substantial body of literature on archetypal emotions and speech, dating back to the 1930s [31]-[33]. A number of key review articles summarize the main findings, notably Frick [34], Scherer [15], and Murray and Arnott [35]. A second body of literature deals with other emotional states, particularly those which tend to be described as attitudes. Key descriptions are contained in Schubiger [36], Crystal [37], [38], and O'Connor and Arnold [39]. The material is very diverse methodologically. The relevant speech variables are sometimes measured instrumentally, but they are often described in impressionistic terms, which may be highly subjective. A relatively small number of studies leads directly to possible implementations.

Different methodologies also tend to be associated with different emotional domains. By and large experimentalists have focused on archetypal emotions. The ones most studied are anger, happiness/joy, sadness, fear, and disgust; also studied experimentally are surprise/astonishment, grief/sorrow, affection/tenderness, and sarcasm/irony [35]. In contrast, linguists have used a much wider range of labels to describe emotional states or attitudes. Two studies by Schubiger [36] and O'Connor and Arnold [39], for example, used nearly 300 labels between them. These cover states such as "abrupt, accusing, affable, affected, affectionate, aggressive, agreeable, airy, amused, angry, animated, annoyed, antagonistic, apologetic, appealing, appreciative, apprehensive, approving, argumentative, arrogant, authoritative, awed..." Arguments tend to be based on linguists' intuitions and illustrated by examples of particular intonational patterns across phrases or sentences which convey a certain kind of feeling.

The aim of this section is to draw together that material in a way that is reasonably systematic to provide a background against which it is possible to make informed judgments about the features an emotion detection system might use.

Descriptive Frameworks for Speech

Discussions of emotion and speech depend on concepts that are not necessarily well known outside the speech

community. This subsection introduces them and then outlines the kinds of relationship between speech and emotion that have been considered.

Speech Production: Basic Terms and Concepts

The main energy source in speech is vibration of the vocal cords. At any given time, the rate at which vocal cords vibrate determines the fundamental frequency of the acoustic signal, usually abbreviated to F0. F0 corresponds (with some qualifications) to perceived voice pitch. Vocal cord vibration generates a spectrum of harmonics, which is selectively filtered as it passes through the mouth and nose, producing the complex time-varying spectra from which words can be identified. Variations in voice pitch and intensity may also have a linguistic function. The patterns of pitch movement which constitute intonation mark linguistic boundaries and signal functions such as questioning. Linked variation in pitch and intensity mark words as stressed or unstressed. The term prosody refers to the whole class of variations in voice pitch and intensity that have linguistic functions.

Speech presents two broad types of information. It carries linguistic information insofar as it identifies qualitative targets that the speaker has attained (or approximated) in a configuration that conforms to the rules of language. Paralinguistic information is carried by allowed variations in the way that qualitative linguistic targets are realized. These include variations in pitch and intensity having no linguistic function and voice quality, related to spectral properties that aren't relevant to word identity.

The boundary between those streams is a matter of controversy. Linguists assume that there are qualitative targets that are understood intuitively by users, but not yet fully explicated, and that they actually account for a good deal of variation that is (mistakenly) classed as paralinguistic. In particular, they tend to look for targets of that kind which underlie the expression of emotion. In contrast, biologists and psychologists tend to assume that the relevant information is defined by continuous variables, which carry paralinguistic information. That parallels the more general questions raised in an earlier section about the roles of qualitative targets (linguistic) and manner of production (paralinguistic).

Table 7. Table of Speech and Emotion (b).

		Fear	Grief	Surprise/Astonishment
Acoustic	Pitch	Increase in mean F0 [157], [171], [175], range F0 [158], [175], perturbation [158], [176], variability F0 movement [158]	Very low range [32], low median [32], raised mean F0 [177], slow change [32]	Wide range [166], median normal or higher [168]
	Intensity	Normal		
	Duration	Increased rate [162], [171], reduced rate [176]	Slow—due to high rate of pause to phonation time, longer vowels and consonants [165]	Tempo normal [168], tempo restrained [166]
	Spectral	Increase in high-frequency energy		
Contour		Disintegration of pattern and great number of changes in direction of pitch [32]	Long sustained falling intonation throughout each phrase [49]	Sudden glide up to a high level within the stressed syllables, then falls to mid-level or lower level in last syllable [166]
Tone Based				Fall rise nuclear tone with falling head (in questions) [39], high fall preceded by rising head (in interjections) [39], high rise tone [38]
Voice Quality		Tense [15]	Whisper [49]	Breathy [166]
Other		Precise articulation of vowel/consonant [165], voicing irregularity due to disturbed respiratory pattern [165]	Voicing irregularities [49]	

Targets and Manners Associated with Emotional Expression

Four broad types of speech variable have been related to expression of emotional states, divided in two groups, in line with a controversy over the roles of linguistic and paralinguistic information.

Tone Types as Targets: An explicitly linguistic tradition focuses on what we will call the tone-based level of description. It tends to be rooted in the British approach to intonation, which describes prosody in terms of intonational phrases or tone groups. Each tone group contains a prominent or nuclear tone (a rising or falling movement or combination or level tone usually on the last stressed syllable of the group). The part that leads up to this nuclear tone is called the head or pretonic, and the part following it is the tail.

Studies in that tradition often associate different types of tones and heads with different emotions or attitudes. A relation between tone shape (e.g., rising or falling) and emotion is claimed, as is a relation between the phonetic realization of a tone (e.g., high rise versus low rise) and emotion. Heads also can take different shapes (e.g., level, falling), and it is claimed that the head or pretonic shape (in conjunction with the tone shape and realization) can express different emotions. The emotions listed cover a wide spectrum. Sometimes particular patterns are listed simply as emotional or nonemotional, sometimes very specific labels are given (e.g., surprise, warmth) though

Table 7. Table of Speech and Emotion (c).

		Excitement	Warmth	
Acoustic	Pitch	Wide range [38]		
	Intensity	High [38]		
	Duration	Fast [38]		
	Spectral			
Contour				
Tone Based		Falling and rise-fall tones [38]	Wide ascending and descending heads [179]	
Voice Quality				
Other				

Table 7. Table of Speech and Emotion (d).

		Sarcasm/Irony	Boredom	Anxiety/Worry
Acoustic	Pitch		Decrease in mean F0 [32], [156]	Increased in mean F0 [159], [180]
	Intensity		Decreased [156], [163]	
	Duration	Restrained tempo [166]	Increased rate [49], [177], decreased rate [163], [170]	
	Spectral			
Contour		Stressed syllables glide to low level in wide arc [166]		
Tone Based		Low rise-fall tone preceded by rising glissando pretonic [178], level nuclear tone [38]	Level tone [38]	
Voice Quality		Tense articulation leading to grumbling [166], creaky phonation [164]		
Other				

these may be qualified with a comment that the specific label depends on the kinetic accompaniment [38] or on a particular linguistic or grammatical context, e.g., statement, wh-question, yes/no question, command, or interjection [39].

Pitch Contours as Targets: An alternative approach describes pitch variation in terms of geometric patterns that are usually described as pitch contours. These are typically studied in experiments where listeners are presented with different contour types and are asked to indicate what emotion they express, for example, on bipolar scales [41], [42]. Contour types sometimes relate to categories in a phonological system [42], but are sometimes unmotivated by any systematic approach to intonation [41]. Hence there is some ambivalence about their relationship to the linguistic/paralinguistic distinction.

There is speculation in the literature that pitch contour type may be more related to attitude than to archetypal emotions. Scherer et al. [43] suggested that continuous variables might reflect states of the speaker related to physiological arousal, while the more linguistic variables such as contour tended to signal attitudes with a greater cognitive component. In a subsequent study, however, where listeners judged the relation of contour type to two scales, a cognitive-attitude scale and an arousal scale [42], they showed that contour type was related to arousal states. Other studies also indicate the relation of contour type to a wide range of emotional states, from archetypal emotions to attitudes.

Manner of Realization—Continuous Acoustic Measures: Many experimental studies focus on measuring continuous acoustic variables and their correlation with specific emotions (particularly full-blown emotions). It is usually assumed that these measures tap paralinguistic properties of speech. It is reasonably clear that a number of continuous acoustic variables are relevant: pitch (F0 height and range), duration, intensity, and spectral makeup. Studies in this mode sometimes manipulate speech instrumentally to separate out single acoustic parameters [40]. Listeners are then tested to see if they can identify what emotion is being expressed, for example, by F0 alone. These experiments suggest that at least some emotional signals are carried paralinguistically.

Manner of Realization—Voice Quality: The fourth level of speech related to the expression of emotion is voice quality. This level is discussed by researchers working within both the experimental and the linguistic tradition. Many describe voice quality auditorily. Terms often used are tense, harsh, and breathy. There is also research, however, which suggests how auditory qualities may map on to spectral patterns [7], [8]. Voice quality seems to be described most regularly with reference to full-blown emotions.

Clearly there are relationships among the levels described above. For example, continuous spectral variables relate to voice quality, and the pitch contours described in the experiments must relate to the tune patterns arising from different heads and tones. But links are rarely made in the literature.

Speech and the Physiology of Emotion

Various physiological changes associated with emotion might be expected to affect speech quite directly, and attempts have been made to infer vocal signs of emotion on that basis [187]. Physiological arousal in general might be expected to affect measures related to effort, such as intensity, mean voice pitch, and speech rate. The tremor associated with fear and anger would be expected to

Table 7. Table of Speech and Emotion (e).

		Affection/Tenderness	Coolness/Hostility	Puzzlement
Acoustic	Pitch	Higher mean [166], lower mean [156], narrow range [166]		High mean [38], wide range [38]
	Intensity	Reduced [166]		Low [38]
	Duration	Slow rate [156]		Slow [38]
	Spectral			
Contour		Slightly descending melody [166], steady and slightly upward inflection [156]		
Tone Based			Low falling nuclear tone [39], high head followed by rise-fall nuclear tone [39]	Rising tones [38]
Voice Quality		A little nasal articulation [166]		
Other		Audible off-glide in long stressed syllables [166]		

produce corresponding oscillations in pitch. It has been suggested that unpleasantness is likely to lead to tensing of the vocal tract walls and hence to alter spectral balance.

Effects of these types are bound to play a part in emotional speech, particularly when the emotion is extreme. However, empirical tests show mixed success for predictions based on them [8]. That is also to be expected, since there are social and biological reasons to ensure that emotion rarely becomes extreme enough to distort speech and that there are signs which can be given before it does.

The neural basis of emotional speech has not been investigated in depth. Some intriguing experiments using electrical stimulation suggest relations between the emotional content of speech and sites on the right side of the cortex, analogous to those on the left for meaning (Wernicke's area) [83], [84]. The amygdala have also been implicated in a recognition code for fearful expressions [92].

Recurring Problems

A Descriptive Framework for Emotional States: The first recurring problem is summarized by Ladd et al., referring to their own study.

> Perhaps the most important weakness of this study, and indeed of the whole general area of research, is the absence of a widely accepted taxonomy of emotion and attitude. Not only does this make it difficult to state hypotheses and predictions clearly, but (on a more practical level) it makes it difficult to select appropriate labels in designing rating forms [42, p. 442].

Couper-Kuhlen [44] makes a similar point.

The clearest symptom of the problem is the multiplicity of categories that linguistic studies use to describe emotional states. The fact that almost no two studies consider the same categories makes integration extremely difficult. Multiplicity is less of a problem on the experimental side because of its emphasis on the archetypal emotions, but coherence is only achieved by a drastic restriction of scope. Unless the linguists are totally misguided, speech signals a range of emotion-related phenomena that is much wider and richer than the archetypal emotions. Ideas considered earlier suggest how those problems could be addressed, i.e., by developing ways of covering a broad domain coarsely, using a relatively small number of categories or dimensions. Indications that that approach may be useful are considered below, after reviewing relevant evidence.

Table 8. Emotion Attributions and Features of Deafened People's Speech.

Response	Speech Factors
Judged stability	Relatively slow change in the lower spectrum
Judged poise	Narrow variation in F0 accompanied by wide variation in intensity
Judged warmth	Predominance of relatively simple tunes, change occurring in the mid-spectrum rather than at extremes; low level of consonant errors
Competence	Pattern of changes in the intensity contour

Relations between Types of Information Source: Linguistic and paralinguistic information are logically linked, but there is a tendency to ignore the relationship between them. Ladd [45] has recently discussed the issue and highlighted evidence that there are significant relationships between the two types of information. He cites two experiments reported by Scherer et al. [43]. In the first, judges agreed on the emotions of question utterances with the words removed, i.e., based on the pitch contour alone. The outcome demonstrated that some of the emotional content of an utterance is indeed nonphonological. The second experiment then presented the utterances without removing the words. Categorical analysis of the pitch movements involved revealed that judgments were affected by linguistic function. For example, yes-no questions with a final fall were rated strongly challenging; rising yes-no questions were rated high on a scale of agreeableness and politeness, while falling yes-no questions were rated low on the same scale.

In terms of the source types introduced in the last section, it is implied that paralinguistic cues are at least partly contextual—their meaning cannot be determined without reference to other features.

Ecological Validity: It is a feature of the literature that it tends to deal with highly artificial material. Most studies use actors' voices rather than examples of naturally occurring emotions. Data sets are often limited to short, isolated utterances (often sentence length). One reason for the use of artificial data is presumably that it is easier to collect than genuine emotional data—particularly if the aim is to study archetypal emotions. More specifically, experimental studies have been driven by concern to avoid presenting verbal cues. That has led them to select or to modify material in a variety of ways. Key examples are as follows.

▲ *Meaningless Content*: Speakers express emotions while reading semantically neutral material, and listeners are asked to identify the intended emotion. For example, Davitz and Davitz [46] had subjects read sections of the alphabet while expressing a series of emotions. Listeners could identify intended emotion in majority of cases.

▲ *Constant Content*: Comparison of the same sentence given by speakers expressing different emotions. For example, Fairbanks [32], [47]-[49] recorded amateur actors reading a semantically neutral phrase in a variety of emotions. Listeners were able to identify emotions correctly most of the time.

▲ *Obscuring Content*: Either by measuring only specific nonverbal properties or by electronically filtering the

Some speech attributes seem to be associated with general characteristics of emotion, rather than with individual categories.

speech itself to obliterate word recognition. For example, Starkweather [50] recorded speech from vocal role-playing sessions and analyzed listeners' perception of the aggressiveness/pleasantness of the speech under three types of presentation—as a normal recording, as a written transcript, and as a filtered content-free recording. Emotion was better perceived from the filtered content-free speech than from the transcript.

These procedures aim to exercise control over the linguistic content, i.e., hold it steady so that it can be separated off from the paralinguistic. The strategy would be appropriate if the levels were independent. As was noted, however, there is evidence that paralinguistic information can function contextually, interacting with linguistic information. Thus, the control strategy risks circularity; a tacit model governs data collection, in a way that precludes finding evidence that could expose problems inherent in the model. Nevertheless, methodologies that are oriented towards ecological validity have begun to emerge. An example is Roach's [51] use of material recorded from radio and television shows: emotional labels are attached on the basis of listener response.

Although the literature on emotion and speech lacks integration, it does offer a substantial body of data. The next section attempts to pull this data together.

Speech Parameters and Specific Emotions

Table 6 is a summary of relationships between emotion and speech parameters from a review by Murray and Arnott [35]. A similar level of description, with slightly different content, is given by Cahn [202]. Those summaries reflect their function, which is to allow the synthesis of voices showing archetypal emotions. For that task it is not necessary to consider either other emotional states or variant ways of expressing archetypal emotions. The same does not apply in research that aims to deal with signs of emotion generated by human speakers, because they may move through a wider range of states and use variant forms of expression. In that context, a fuller overview is useful. Table 7 sets out a summary that covers most of the available material on the speech characteristics of specific emotions.

The emotional states are those which are most commonly mentioned in the speech literature. They are ordered to reflect earlier sections. The first five are standard archetypal emotions. They are ordered in terms of emotional orientation as measured in the BEEVer study [197], so that adjacent columns are relatively close in activation/evaluation space. Grief is placed in the sixth column to reflect its relationship to sadness. The next four columns deal with terms that are not usually cited as archetypes, but that a reasonable proportion of BEEVer subjects included as part of a basic emotion vocabulary. The remaining items shade into the domain of attitude.

The description of each state reflects the division of speech variables set out above (continuous acoustic, pitch contour, tone based, voice quality). A fifth category called Other is included to cover variables which do not fall easily into any of the previous four. The data is taken from a range of studies, which are referenced in Table 7. When there are blanks under one of the speech categories for a particular emotion, this means that no relevant studies have been found. Descriptions under the category Acoustic generally mean that the emotion has been shown to contrast with neutral speech. Speech attributes given in bold typeface mean that these seem to be reliable indicators of the emotion, i.e., occur across a number of studies and the data is substantial. Table 7 highlights four broad points:

▲ The first point is that a good deal is known about the speech correlates of the archetypal emotions of anger, happiness, sadness, and fear. The speech measures which seem to be reliable indicators of the primary emotions are the continuous acoustic measures, particularly pitch-related measures (range, mean, median, variability), intensity, and duration.

▲ The second point is that our knowledge is, nevertheless, patchy and inconclusive. There are a several pointers to this. First, even within the archetypal emotions, there are contradictory reports. For example, there is disagreement on duration aspects of anger, happiness and fear—some report longer duration, some report faster speech rate, and some report slower speech rate. Second, the large gaps under some headings in the table indicate incomplete knowledge. Third, our knowledge at the level of voice quality is noticeably incomplete. Attributes of voice quality are often mentioned, but they are mostly auditorily judged; only a few studies tie voice quality to instrumental measures [7], [8].

▲ The third point is the lack of integration across the paralinguistic (as represented by the continuous acoustic level) and the linguistic (as represented by the tone-based level). The evidence indicates that continuous speech attributes are related to the archetypal emotions and that linguistically based attributes are related to nonarchetypal emotions. That may be because archetypal emotions are signaled paralinguistically and others by linguistic signs. Alternatively, it may simply reflect the way that certain methodologies have traditionally been used to study certain kinds of emotional state. In the absence of direct evidence, we do not know.

▲ The fourth point is that some speech attributes seem to be associated with general characteristics of emotion, as discussed previously, rather than with individual categories. The clearest examples involve activation. Positive activation appears to be associated with increased mean and

range of F0, and tense voice quality—consider happiness, fear, anger, and to a lesser extent surprise, excitement and puzzlement. Negative activation appears to be associated with decreased mean and range of F0—consider sadness, grief, and to a lesser extent boredom. Disposition to seek information is also associated with reports of upward movement in the contour and tone-based categories—consider surprise, happiness, and puzzlement. These associations are not conclusive, but they do strengthen the case for one of the strategies considered in the previous section, that is, to explore correspondences between expressive signs and features that are common to a range of emotional states.

Computational Studies of Emotion in Speech

There are relatively few systems which approach the goal of recognizing emotion automatically from a speech input. This section reviews key examples.

Cowie and Douglas-Cowie ASSESS System

ASSESS [1], [7] is a system which goes part way towards a computational analysis. Automatic analysis routines generate a highly simplified core representation of the speech signal based on a few landmarks—peaks and troughs in the profiles of pitch and intensity and boundaries of pauses and fricative bursts. These landmarks can be defined in terms of a few measures. Those measures are then summarized in a standard set of statistics. The result is an automatically generated description of central tendency, spread and centiles for frequency, intensity, and spectral properties.

That kind of feature extraction represents a natural first stage for emotion recognition, but in fact ASSESS has not generally been used in that way. Instead the measures described above have been used to test for differences between speech styles, many of them at least indirectly related to emotion. The results indicate the kinds of discrimination that this type of representation could support.

A precursor to ASSESS was applied to speech produced by deafened adults [52]. One of the problems they face is that hearers attribute peculiarities in their speech to emotion-related speaker characteristics. These evaluative reactions were probed in a questionnaire study, and the programs were used to elicit the relevant speech variables. Table 8 summarizes correlations between emotion attributions and speech features. They suggest that ASSESS-type measures are related to judged emotionality, but also underline the point that the attribution of emotion is dogged by systematic ambiguity.

Simulation

In a later study, reading passages were used to suggest four archetypal emotions: fear, anger, sadness, and happiness [53]. All were compared to an emotionally neutral passage, and all passages were of comparable lengths. Speakers were 40 volunteers from the Belfast area, 20 male and 20 female, between the ages of 18 and 69. There was a broad distribution of social status, and accents represented a range of local types. Subjects familiarized themselves with the passages first and then read them aloud using the emotional expression they felt was appropriate. Recordings were analyzed using ASSESS. Table 9 summarizes the measures that distinguish the emotionally marked passages from the neutral passage.

Figure 4 presents traces from an individual speaker—arbitrarily chosen from the group studied by McGilloway [54]—and shows how they relate to the kinds of features suggested by ASSESS analysis. Figure 4(a)-(e) summarizes the output of initial processing on each of five signals—one neutral and four expressing specified emotions (anger, fear, happiness and sadness). Time, in milliseconds, is on the horizontal axis. The heavy lines in each figure show signal intensity (referred to the left-hand scale, in decibels), the light lines represent pitch (referred to the left-hand scale, in hertz). Time scale (on the horizontal axis, in milliseconds) is adjusted to let the whole trace appear on the figure. The patterns are summaries in that inflections and silences have already been identified from the raw input, and the overall contours are represented by a series of straight lines (or gaps) between the resulting points. Several spectrum-like representations have also been computed, but found to contribute relatively little.

It is not self-evident from inspection that the contours differ systematically, but analysis indicates that they do. Figure 4(f) shows how the person's speech re-

Table 9. Distinctions between Emotional and Neutral Passages Found by ASSESS.

	Spectrum		Pitch Movement		Intensity		Pausing	
	Midpoint and Slope		Range	Timing	Marking	Duration	Total	Variability
Afraid					+	+		
Angry		+	+		+	+		
Happy			+	+	+		+	+
Sad				+		+	+	

lates to the general distinctions found in the whole subject group (n=40). Each caption on the left-hand side refers to an output feature whose value in one or more emotional traces is significantly different from its value in the neutral passage. The features are selected from a much larger number that meet that basic criterion. Selection is geared to a) avoiding redundancy, b) representing the main logically distinct areas where differences occur, and c) achieving some formal consistency (e.g., using centile measures to describe central tendency and spread). The graph shows the fit of the speaker's data to a template based on the overall analysis. A bar is present if the feature in question differentiates the emotion from neutral in the analysis. Its direction is positive if the difference is in the direction indicated by the analysis. Its length is proportional to the difference between the feature value for the emotion in question and the same value for the neutral signal, relative to the standard deviation of the values for all expression classes (four emotions plus neutral) on that feature.

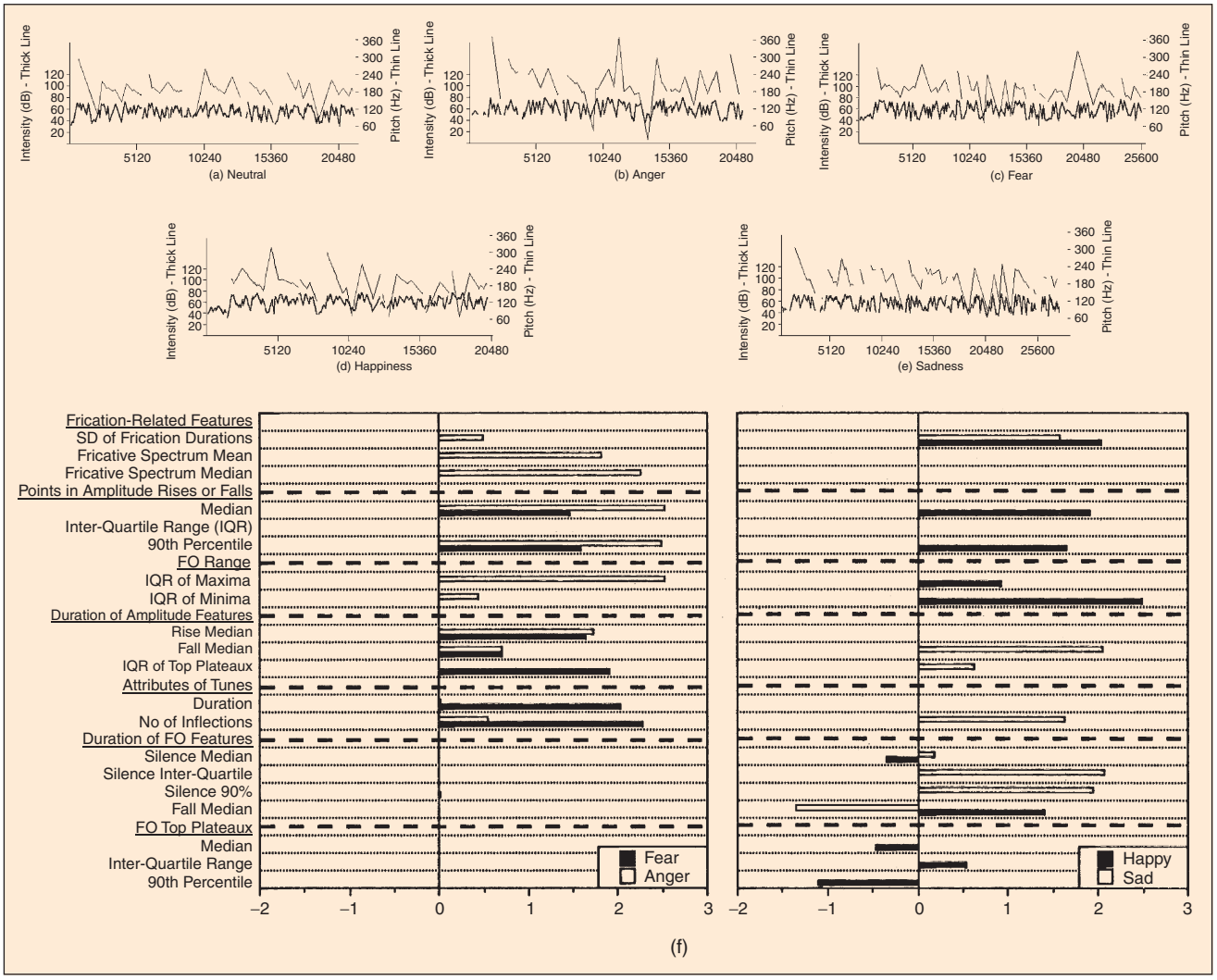
The main points to be made from Fig. 4(f) are that the kind of analysis embodied in ASSESS generates a range of features that are relevant to discriminating emotion from neutral speech and that different emotions appear to show different profiles. It remains to be seen how reliably individual signals can be assigned to particular emotion categories, but there are grounds for modest optimism.

Discriminant Analysis

A recent application of the system illustrates the next natural step towards automatic discrimination [9]. Discriminant analysis was used to construct functions which partition speech samples into types associated with different communicative functions, e.g., opening or closing the interaction, conducting business. It is an interesting question how closely that kind of functional difference relates to emotion.

Banse and Scherer' s System

Scherer's group has a long record of research on vocal signs of emotion. A key recent paper [8] extracts a systematic battery of measurements from test utterances. The measures fall into four main blocks, reflecting the consensus of



▲ *4. (a)-(e) Output of initial processing on each of five speech passages, (f) How a person's speech relates to the general distinctions found in the whole subject group.*

research concerned with the continuous acoustic level (see Table 10). The emotions considered were hot anger, cold anger, panic fear, anxiety, desperation, sadness, elation, happiness, interest, boredom, shame, pride, disgust, and contempt. Discriminant analysis was then used to construct functions which partition speech samples into types associated with different types of expression. Classification by discriminant functions was generally of the order of 50% correct—which was broadly comparable with the performance of human judges.

It is natural to take the techniques described by Banse and Scherer as a baseline for emotion detection from speech. They show that automatic detection is a real possibility. The question then is where substantial improvements might be made. That theme is taken up below.

Automatic Extraction of Phonetic Variables

One of the major tasks facing automatic emotion detection is automatic recovery of relevant features. Large parts of the literature described above consider features which can be identified by human observers, but which have no simple correlate in the acoustic signal.

ASSESS reflects one approach. It uses features which can be derived in a relatively direct manner from the acoustic signal—though even in that case the processing involved is far from trivial and it depends on human intervention in the case of noisy signals. However, modern signal processing techniques mean that a far wider range of features, in principle, could be explored. This section begins with a case study which illustrates some relevant issues and then considers the main feature types that are of interest in turn.

Voice Stress: A Case Study

Voice level is one of the intuitive indicators of emotion. Banse and Scherer measure it in what is the most obvious way, as a direct function of microphone voltage. However, simple relationships between voltage and voice level only exist under very special circumstances. Normally, microphone voltage depends critically on the distance between the speaker and the microphone, on the direction in which the speaker is turned, and on reflection and absorption of sound in the environment. Humans provide an existence proof that it is possible to compensate for these effects—they can usually tell the difference between a person whispering nearby and a person shouting far off.

A report by Izzo [55] considers the kinds of solution to this problem that modern engineering makes available. The context is speaker stress, which has direct applications, but what is most relevant to this context is the fact that the problem includes distinguishing loudness-related varieties—soft, neutral, clear, and loud. It uses speech from databases (SUSAS and TIMIT) which consist of short utterances labeled in detail. The following indicators are considered.

Pitch is shown to be a statistical indicator of some speech types (e.g., clear and soft). The duration of speech sounds can be established because of the labeling, and it is indicative of speech type—particularly the duration of semivowels. Intensity per se is subject to the confounding factors which have been mentioned above, but the distribution of energy is also an indicator of speech type—for instance, energy shifts towards vowels and away from consonants in loud speech.

It is well known that the spectral distribution of energy varies with speech effort—effortful speech tends to contain relatively greater energy in low- and mid-spectral bands. That kind of relationship is exploited both in the ASSESS family [52] and by Banse and Scherer. Izzo examines a number of ways in which the approach can be refined. Wavelet transforms provide a more flexible method of energy decomposition than the Fourier-based techniques used in earlier work. Discrimination is increased by distinguishing the spectra associated with different speech sounds. Time variation in the energy distribution is also more revealing than static slices or averages. Standard techniques allow the cross section of the vocal tract to be estimated from particular speech sounds, and they show that speech level affects the region in which greatest movement occurs during production of a vowel sound.

The key message of the study is that intervening variables are central to the area. Voice level itself is an intervening variable—it is an indicator of emotion, but extracting it is a substantial task. Because of their potential relationship to the biology of emotion, intervening variables which refer to physiological states—such as vocal tract configuration—are particularly interesting, and there are techniques which allow them to be recovered. The use of information about speech sounds highlights the relevance of what may be called intervening covariables. Voice level may be a paralinguistic feature, but it is not necessarily optimal to ignore linguistic issues

Table 10. Banse and Scherer's Measures Concerning Continuous Acoustic Level.

Fundamental frequency	Mean F0	Standard Deviation of F0	25th and 75th Percentiles of F0
Energy	Mean of log-transformed microphone voltage		
Speech rate	Duration of articulation periods		Duration of voiced periods
Spectral measures	Long-term average spectra of voiced and unvoiced parts of utterances		

(such as phoneme identity) in the process of recovering it. Stress may also be considered as an intervening variable—a feature which distinguishes certain emotional states from others.

The issue of intervening variables is of particular interest for neural networks. On the one hand, it is an attraction of neural nets that they have the potential to allow evidence to drive the emergence of suitable intervening structures. On the other, it is a danger that they may generate weighting patterns which work—particularly in a restricted domain—but which can neither be understood nor extended. Hybrid structures offer the prospect of addressing those issues [56].

Relevant Feature Types

This section sets out to summarize the kinds of intervening variables that it makes sense to consider extracting from the raw input and the techniques that are currently available.

Voice Level: This was considered in the previous section.

Voice Pitch: Voice pitch is certainly a key parameter in the detection of emotion. It is usually equated with F0. Extracting F0 from recordings is a difficult problem, particularly if recording quality is not ideal. It involves several subproblems, such as detecting the presence of voicing, the glottal closure instant [56], the harmonic structure in a brief episode [58], short-term pitch instabilities (jitter and vibrato) [59], and fitting continuous pitch contours to instantaneous data points.

Phrase, Word, Phoneme and Feature Boundaries: Detecting boundaries is a major, but difficult, issue in speech processing. That is why recognition of connected speech lags far behind recognition of discrete words. The issue arises at different levels.

▲ *Phrase/Pause Boundaries*: The highest level boundary that is likely to be relevant is between a vocal phrase and a pause. Quite sophisticated techniques are available to locate pauses [60]. In [7] a method based on combining several types of evidence is used, and it is reasonably successful. However, the process depends on empirically chosen parameters, and it would be much better to have them set by a learning algorithm—or better still, by a context-sensitive process. As noted above, pause length and variability do seem to be emotionally diagnostic.

Table 11. Duration Features and Emotions (ms).

	Rises	Falls	Tunes	Plateau
	Median	Median	Median	IQR
Fear	82.35	84.8	1265	10.8
Anger	81.66	80.5	1252	10.2
Happiness	78.03	77.4	1404	8.2
Neutral	78.50	77.2	1452	8.4
Sadness	77.28	81.4	1179	11.0

▲ *Word Boundaries*: Speech rate is emotionally diagnostic, and the obvious way to describe it is in words per minute, which depends on recovering word boundaries. That turns out to be an extremely difficult task, and probably the best solution is to look for other measures of speech rate, which lend themselves better to automatic extraction. Finding syllable nuclei is a promising option [61], [62].

▲ *Phoneme Boundaries*: The report by Izzo indicates that good use can be made of information about phonemes if they can be identified. That directs attention to a large literature on phoneme recognition [63]-[65].

▲ *Feature Boundaries*: Some features, such as fricative bursts, are easier to detect than phonemes as such and they appear to be emotionally diagnostic.

Voice Quality: A wide range of phonetic variables contribute to the subjective impression of voice quality [66]. The simplest approach to characterizing it is based on spectral properties [67]. The report by Izzo reflects that tradition. A second uses inverse filtering aimed at recovering the glottal waveform (another task where neural net techniques can be used to set key parameters [68]). Voice quality measures, which have been directly related to emotion, include open-to-closed ratio of the vocal cords, jitter, harmonics-to-noise ratio, and spectral energy distribution [69].

Temporal Structure: This heading refers to measures at the pitch contour level and related structures in the intensity domain. ASSESS contains several relevant types of measure. The pitch contour is divided into simple movements: rises, falls, and level stretches (see Table 11). Describing pitch movement in those terms appears to have some advantages in the description of emotion over first-order descriptions (mean, standard deviation, etc). The intensity contour is treated in a similar way, and again, descriptions based on intensity movements seem to improve emotion-related discriminations.

ASSESS also incorporates simple measures of tune shape. Portions of pitch contour between adjacent pauses are described in terms of overall slope and curvature (by fitting quadratic curves). Research at the University of Pittsburgh, studying mothers' speech to infants, has explored wider range of fitted curves, including exponential, Gaussian, and sinusoidal. The fits of linear, power, and exponential curves contributed to discriminant functions distinguishing between two types of utterance with emotional overtones—expressing approval and seeking attention [186]. An adjusted classification allowed discrimination between these and a third category (giving comfort). It categorized curves as rising or falling and used a visual judgment of wave-like fluctuation [201].

A natural extension to this kind of description is to consider rhythm. Rhythm is known to be an important aspect of speech [70], [71], but few measures are available. Progress has been made on a simple aspect of rhythm, the alternation between speech and silence.

Stress has been shown to have a dual effect on pause patterns in discourse [181]. Stressed subjects show shortened switching pauses (i.e., pauses before they begin a turn) and lengthened internal pauses (i.e., pauses during the course of a turn). Following that observation, the Pittsburgh group has shown that depressed mothers use switching pauses which are abnormal in several respects—lengthened and highly variable, both in absolute terms and relative to mean length [188]. ASSESS contains procedures designed to measure regularity of emphasis, but they are not satisfactory.

Linguistically Determined Properties: There is a fundamental reason for considering linguistic content in connection with the detection of emotion. On a surface level, it is easy to confound features which signal emotion and emotion-related states with features which are determined by linguistic rules. The best known example involves questions, which give rise to distinctive pitch contours that could easily be taken as evidence of emotionality if linguistic context is ignored. Some work has been done on the rules for drawing the distinction [72]. Other linguistic contexts which give rise to distinctive pitch contours are turn taking [73], topic introduction [74], and listing.

It is worth noting that these are contexts that are likely to be quite common in foreseeable interactions with speech competent computers: systematically misinterpreting them as evidence of emotionality would be a nontrivial problem. The only obvious way to avoid confounding in these contexts is to incorporate intervening variables which specify the default linguistic structure and allow the observed speech pattern to be compared with it.

Natural Directions for Research

This section attempts to identify the natural directions for research by taking Banse and Scherer [8] as a point of reference, identifying where their approach is incomplete or questionable and considering how it could be taken forward.

Perhaps the most obvious priority is to extend the range of speech material from which classifications can be made. Ideally speech samples should:

▲ be natural rather than read by actors (who presumably tend to maximize the distinctiveness of emotional expression);

▲ have verbal content varying as it naturally would rather than held constant (so that potential confounding effects have to be confronted);

▲ include rather than exclude challenging types of linguistically determined intonation;

▲ be drawn from a genuine range of speakers, in terms of sex, age, and social background;

▲ use a range of languages.

A second priority is to extend the range of evidence that may be used in classification, and examine what, if anything, they contribute, particularly in less constrained tasks. Relevant types include:

Relationships to nonemotional states are also a very real issue. A husky voice may reflect either passion or a sore throat.

▲ derived speech parameters of the kinds considered in "Relevant Feature Types";

▲ linguistic information of the kinds considered in "Relevant Feature Types"

▲ nonspeech context, particularly facial signals, considered in the next section.

Extending beyond speech information is not gratuitous. It reflects a point which has been made repeatedly, that cues to emotion may be at least partly contextual—i.e., their meaning cannot actually be determined without reference to other features of the situation. It has been shown that that is the case for at least some paralinguistic cues. As a result, considering speech without reference to other sources risks misrepresenting the kind of information that it provides.

The third priority is to extend the range of responses. Increasing the range of emotion terms considered is one aspect of the issue, but considering how that might be done quickly indicates that it entails deeper changes, since it makes no sense to treat hundreds of emotion terms as independent entities. Relationships among them need to be considered systematically, in several senses.

▲ They may be located in dimensions of the kinds considered in "A Descriptive Framework for Emotional States"; it then becomes a priority to consider mappings between those dimensions and dimensions of speech variation;

▲ They may be related to intermediate features, which characterize a range of possible emotions (e.g., stressed, positive); it then becomes a priority to consider mappings between those intermediate features and dimensions of variation in speech.

▲ They may be linked to possible actions which apply under uncertainty, e.g., "ask whether X," "look out for Y," "just get out fast."

Relationships to nonemotional states are also a very real issue. It is important to register that a husky voice may reflect either passion or a sore throat, and that is a problem for models which propose automatic links between voice parameters and emotional attributions.

Once again, the last two points underline the need to consider contextual issues in the use of speech-based information about emotion. It seems highly likely that people are able to do much with a rather narrow paralinguistic channel because evidence from it feeds into knowledge-rich inference processes, and as a result they can make the right attribution for effects that have many potential causes. This may be wrong, but it is plausible enough to suggest that simpler models should not be taken for granted.

Faces and Emotional States

There is a long history of interest in the problem of recognizing emotion from facial expressions [98], influenced by Darwin's pioneering work [76] and extensive studies on face perception during the last 20 years [77]-[79]. Traditional approaches have been supplemented by research on the neural mechanisms involved in the process [81].

Descriptive Frameworks for Emotional Expression in Faces

The salient issues in emotion recognition from faces are parallel in some respects to the issues associated with voices, but divergent in others.

As in speech, a long established tradition attempts to define the facial expression of emotion in terms of qualitative targets—i.e., static positions capable of being displayed in a still photograph. The still image usually captures the apex of the expression, i.e., the instant at which the indicators of emotion are most marked. More recently, emphasis has switched towards descriptions that emphasize gestures, i.e., significant movements of facial features.

Marked contrasts with the speech literature arise from the involvement of different disciplines. In particular, neurophysiology has made more progress towards understanding how humans recognize emotion from faces, and there is a broader base of relevant computational research [81], [192]. Both are linked to the major effort which has been devoted to recognizing individuals from facial information, usually referred to simply as "face recognition" [82].

A final contrast is that in the context of faces, the task has almost always been to classify examples of archetypal emotions. That may well reflect the influence of Ekman and his colleagues, who have argued robustly that the facial expression of emotion is inherently categorical.

There is some literature on nonarchetypal expressions. Intriguingly, Ekman and Friesen [104] have made observations on the way that facial expressions may be modulated (by changing the number of facial areas involved, controlling timing, and adjusting the strength of muscle pull) or falsified (by simulating, neutralizing, or masking one emotion with an expression associated with another). They have also discussed how mixed expressions might reflect the mixed emotions

> that might occur, for example, if Patricia had just approached a reckless driver who had just run over her dog. Saddened at the death of her pet, angry at the driver ... she might blend two feelings in [her] expression. [104, p. 125]

More recently, morphing techniques have been used to probe states that are intermediate between archetypal expressions. They do reveal effects that are consistent with a degree of categorical structure in the domain of facial expression, but they are not particularly large, and there may be alternative ways of explaining them—notably by considering how category terms and facial parameters map onto activation-evaluation space [96].

In practice, though, technical research on the facial expression of emotion has overwhelmingly taken the term emotion in its narrow sense. There seems to be no systems that begin to address the task of reading Patricia's face as she struggles to control her conflicting feelings.

Targets and Gestures Associated with Emotional Expression

Analysis of the emotional expression of a human face requires a number of preprocessing steps which attempt to detect or track the face; locate characteristic facial regions such as eyes, mouth, and nose on it; extract and follow the movement of facial features, such as characteristic points in these regions; or model facial gestures using anatomic information about the face.

Facial features can be viewed [104] as either static (such as skin color), slowly varying (such as permanent wrinkles), or rapidly varying (such as raising of the eyebrows) with respect to time evolution. Detection of the position and shape of the mouth, eyes, particularly eyelids, wrinkles, and extraction of features related to them are the targets of techniques applied to still images of humans. It has been shown by Bassili [108], however, that facial expressions can be more accurately recognized from image sequences than from a single still image. His experiments used point-light conditions, i.e., subjects viewed image sequences in which only white dots on a darkened surface of the face were visible. Expressions were recognized at above chance levels when based on image sequences, whereas only happiness and sadness were recognized at above chance levels when based on still images. Techniques which attempt to identify facial gestures for emotional expression characterization face the problems of locating or extracting the facial regions or features, computing the spatio-temporal motion of the face through optical flow estimation, and introducing geometric or physical muscle models describing the facial structure or gestures.

Most of the above techniques are based on the work of Ekman and Friesen [98], who produced a system for describing "all visually distinguishable facial movements," called the facial action coding system (FACS). FACS is an anatomically oriented coding system, based on the definition of "action units" (AUs) of a face that cause facial movements. Each AU may correspond to several muscles that together generate a certain facial action. As some muscles give rise to more than one action unit, correspondence between action units and muscle units is only approximate. Forty-six AUs were considered responsible for expression control and 12 for gaze direction and orientation. The FACS model has been used to synthesize images of facial expressions; exploration of its use in analysis problems has been a topic of continuous research [85], [91], [99]-[103]. Ekman et al. have also generated, first, a dictionary, called EMFACS, which lists certain key AUs and the actions that can co-occur with them to sig-

nify one of seven archetypal emotions; then they provided a database, called FACSAID, which can serve as a platform for translating FACS scores into emotion measurements [86].

The FACS model has recently inspired the derivation of facial animation and definition parameters in the framework of the ISO MPEG-4 standard [97]. In particular, the facial definition parameter set (FDP) and the facial animation parameter set (FAP) were designed in the MPEG-4 framework to allow the definition of a facial shape and texture, as well as the animation of faces reproducing expressions, emotions, and speech pronunciation. The FAPs are based on the study of minimal facial actions and are closely related to muscle actions. They represent a complete set of basic facial actions, such as squeeze or raise eyebrows, open or close eyelids, and therefore allow the representation of most natural facial expressions. All FAPs involving translational movement are expressed in terms of the facial animation parameter units (FAPU). These units aim at allowing interpretation of the FAPs on any facial model in a consistent way, producing reasonable results in terms of expression and speech pronunciation. The FAPUs are illustrated in Fig. 5(a) and correspond to fractions of distances between some key facial features. FDPs, on the other hand, are used to customize a given face model to a particular face. The FDP set contains a three-dimensional (3-D) mesh (with texture coordinates if texture is used), 3-D feature points, and optionally texture and other characteristics such as hair, glasses, age, gender. The 3-D feature points of the FDP set are shown in Fig. 5(b).

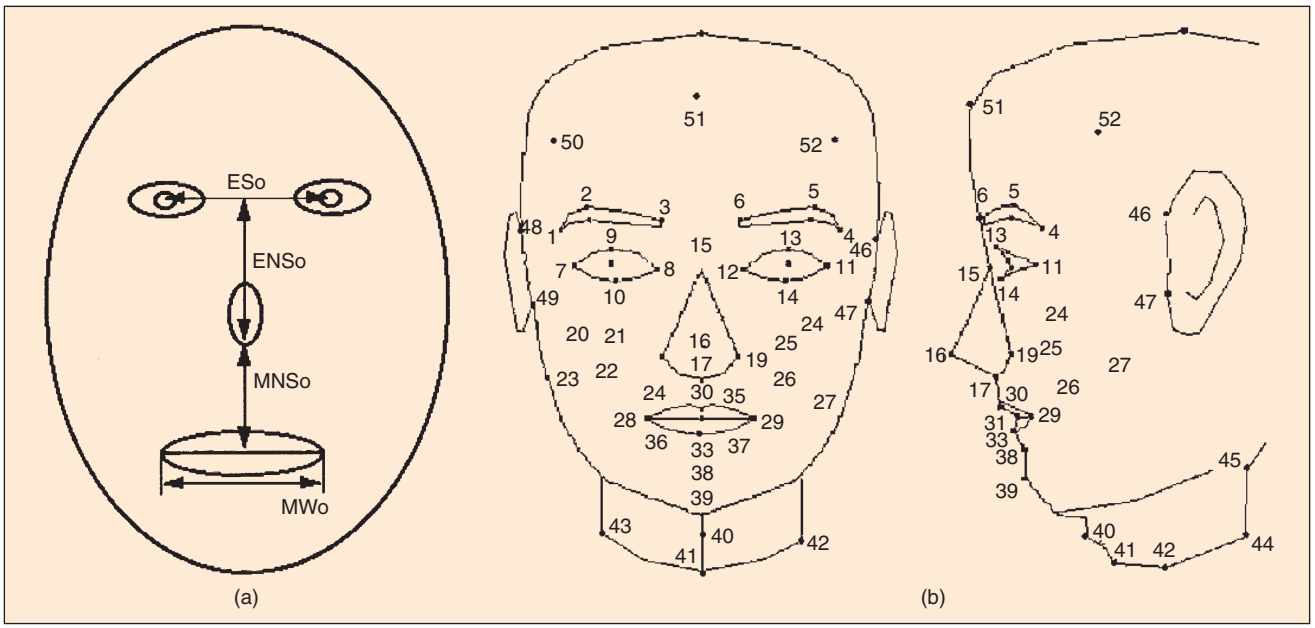
Faces and the Physiology of Emotion

In contrast to speech, neurobiology offers a good deal of information about the recovery of emotion-related information from faces. It is well known that the human right temporal lobe contains a face-sensitive area. Degeneration of this area can lead to prosopagnosia (the inability to recognize faces). More precise localization of face regions in the human brain has become possible through the use of noninvasive brain imaging. This has indicated that there are sites in the occipital and temporal regions, especially those nearest the midline, which are most activated during face processing [87]. Moreover, a number of psychophysical studies have shown that loss of the amygdala, a subcortical nucleus very well connected to many brain areas, causes loss of recognition of anger and fear expressions on human face visual images [75]. Normal volunteers have been shown by fMRI to have increased excitation in the amygdala on viewing expressions of human faces showing mild or strong disgust. This activation even occurs when the subject has no conscious recognition of the face [90]. The whereabouts of sites coding for pleasurable expressions on the human face is unknown at present. However, it would be expected to involve the dopaminergic reward system in the limbic part of the brain as well as in the prefrontal cortex, where face sensitive cells are also observed in monkeys [93].

Overall, the most striking conclusion is that the information relating to facial emotions is processed in rather different sites from information about identity and may itself involve more than one stream.

Computational Studies of Facial Expression Recognition

This section explores methods by which a computer can recover information about emotional state from facial actions-expressions. It emphasizes the task of categorizing active and spontaneous facial expressions so as to extract information about the underlying emotional states.

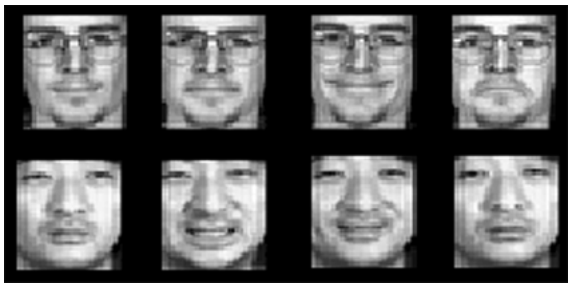


▲ *5. The Facial Animation Parameter Units (FAPUs) and the Facial Definition Parameter (FDP) set defined in the MPEG-4 standard.*

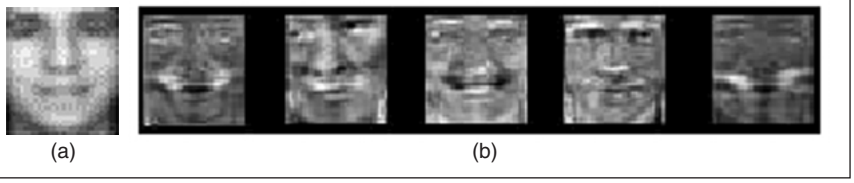
Approaches are divided into two main categories: target oriented and gesture oriented. In target-oriented approaches, recognition of a facial expression is performed using a single image of a face at the apex of the expression. Gesture-oriented approaches extract facial temporal information from a sequence of images in an episode where emotion is expressed, with facial expressions normally lasting between 0.5 and 4 s [79]. Transitional approaches were also developed that use two images, representing a face in its neutral condition and at the apex of the expression.

Face Tracking

In all approaches dealing with the analysis of the emotional expression of faces, the first task is to perform face detection or tracking. Various approaches have been developed for face tracking, without having, however, reached a complete satisfactory solution [129]-[134]. Many tracking systems use color as a clue for detection of facial region mainly due to the fact that the distribution of the chrominance components of a human face are located in a very small region of the color space [135]-[137], but the precision is not very accurate. Other tracking systems use previously given templates, e.g., in the form of gray values, active contours, graphs [138], or wavelets [139], also allowing affine variations of the facial image, but generally are computationally expensive. Image warping based on radial basis functions has been also proposed to account for variations caused by facial expressions [95]. Estimation of the pose of the face is also of particular importance so as to deal with changes in the appearance of faces caused by different viewing angles [140], [141].



▲ *6. Eight sample face images from the CMU dataset of 20 persons showing neutral, angry, happy, and sad facial expressions.*



▲ *7. (a) average face. (b) Weights of five hidden layer neurones of a backpropagation network of size (1295 × 5 × 4) for recognizing the four facial expressions: neutral, angry, happy, and sad.*

Target-Oriented Approaches

Most psychological research on facial expression analysis has been conducted on "mug-shot" pictures that capture the subject's expression at its apex [80]. These pictures allow one to detect the presence of static cues (such as wrinkles) as well as the positions and shapes of facial features. However, extracting the relevant cues from static images has proved difficult, and few facial expression classification techniques based on static images have been successful [105]. An exception [106] used an ensemble of feed-forward neural networks for the categorical perception of static facial emotions. The technique is similar to Eigenface approaches, using seven 32 × 32-pixel blocks taken from the facial regions of interest (both eyes and mouth) and projecting them onto the principal component space generated from randomly located blocks in the image data set. By training each network independently using on-line backpropagation and different input data sets and considering only the network producing the highest output score during recall, an expected generalisation rate of 86% was reached on novel individuals, while humans scored 92% on the same database.

Simulation

Using a modification of a gender recognition system originally proposed by Golomb [107], we show that given prior normalization, neural nets can achieve a basic level of expression classification from still images. A multilayer perceptron network with four output units and one hidden layer was developed to extract a corresponding number of facial expressions from images. The image set used, contained pictures of 20 different males and females. There were 32 different images (maximum size120 × 128) for each person, showing happy, sad, neutral, and angry expressions and looking straight to the camera, left, right, or up. The images were first normalized, resulting in 80 face images of size 35 × 37 (see Fig. 6). Normalization was achieved by detecting the main facial features (eyes, nose and mouth) [94] and by translating, rotating and expanding/shrinking the face around a virtual central (nodal) point. This image set was then split into a training set with nine images and two sets for validation and testing, each containing five images.

Figure 7(b) shows five images of the learned weights of the hidden neurons of the (1295, 54)-MLP network. The third and fourth neurons show similarity to an "eyebrow" detector, which is an important feature for facial expression recognition. Closer inspection of the position of both eyebrows shows a small displacement upwards in the third and downwards in the fourth neuron compared to the average face. These displacements correspond to the happy and angry expressions respectively, which is apparent from the distribution of the neuron's weights. The first and the last

neuron are selective to regions of the mouth and seem to measure the curvature of the lips. This feature is present in most of the images generated at the hidden layer neurons of the network, trained to perform the expression recognition task, suggesting its general importance for face perception. The rotation visible in the second image is caused by the rotation of a face showing an angry expression and displays the perturbation of the network weights caused by an artifact. In Fig. 7(a) an average of all training images is depicted showing a good alignment of the facial contours. The generalization performance, i.e., correct classification of unseen images, was 78%. No doubt numerical performance could be improved. However, it is more important to escape the fundamental artificiality of the task, dealing with extreme cases of a very small number of expressions that vary in many facial regions.

Gesture-Oriented Approaches

Most approaches dealing with facial gestures are based on optical flow estimation. Image gradient, or image filtering, or image correlation are used for estimating optical flow. Gradient algorithms, based on the formulation by Horn and Schunck [109], assume that skin deformations are locally smooth; they face difficulties, however, in highly textured images [101]. Filtering approaches [110] are able to analyze the spatial and temporal frequency content of the image; they require, however, a large number of frames to compute the motion field. Correlation approaches [111], [112] compare a linearly filtered value of each pixel with similarly filtered values of neighboring pixels; they generally are computationally intensive, since some form of exhaustive search is carried out to determine the best estimate of motion. In all cases, the extracted flow patterns can be used by conventional pattern classification techniques as well as neural networks to recognize the corresponding expressions.

Transitional Approaches: Transitional approaches focus on computing motion of either facial muscles or facial features between neutral and apex instances of a face. Mase [101] described two approaches based on muscle motion. In his top-down approach, the facial image is divided into muscle units that correspond to the AUs defined in FACS. Optical flow is computed within rectangles that include these muscle units, which in turn can be related to facial expressions. This approach relies heavily on locating rectangles containing the appropriate muscles, which is a difficult image analysis problem, since muscle units correspond to smooth, featureless surfaces of the face. In his bottom-up approach, the area of the face is tessellated with rectangular regions over which optical flow feature vectors are computed; a 15-dimensional feature space is considered, based on the mean and variance of the optical flow. Recognition of expressions is then based on k-nearest-neighbour voting rule. The results show that even without employing a physical model, optical flow can be used to observe facial motion and track action units.

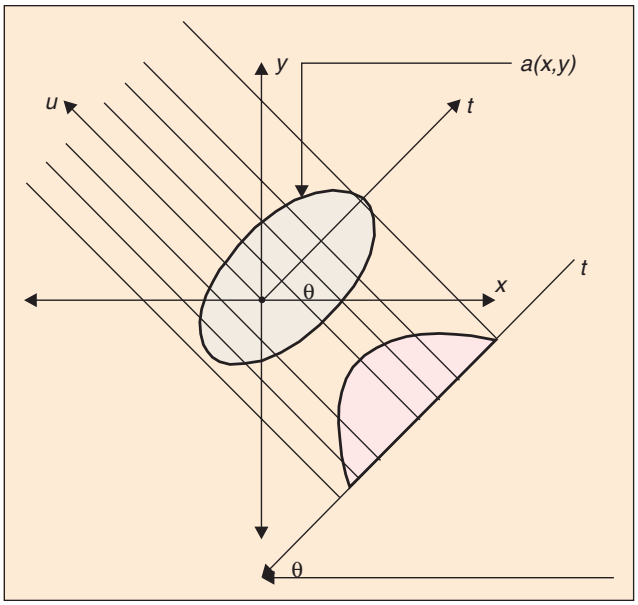
Yacoob and Davis [113], on the other hand, focused on facial edges rather than muscle dynamics, due to the fact that edges and their motion are easier to compute and more stable than surfaces under projection changes. Yacoob unified the facial descriptions proposed by Ekman and Friesen [104] and the motion patterns of expression proposed by Bassili [108], arriving at a dictionary that provides a linguistic, mid-level representation of facial actions, modeling spatio-temporal facial activity. The mid-level representation is computed per frame, thus modeling rapid actions. A rule-based recognition system has been developed, using the descriptions of [104] and [108].

Li et al. [100] described an approach using the FACS model and analyzing facial images for resynthesis purposes. A 3-D mesh was placed on the face, and the depths of points on it were recovered. They derived an algorithm for recovering rigid and nonrigid motion of the face based on two, or more, frames and on six AUs to represent possible facial expressions.

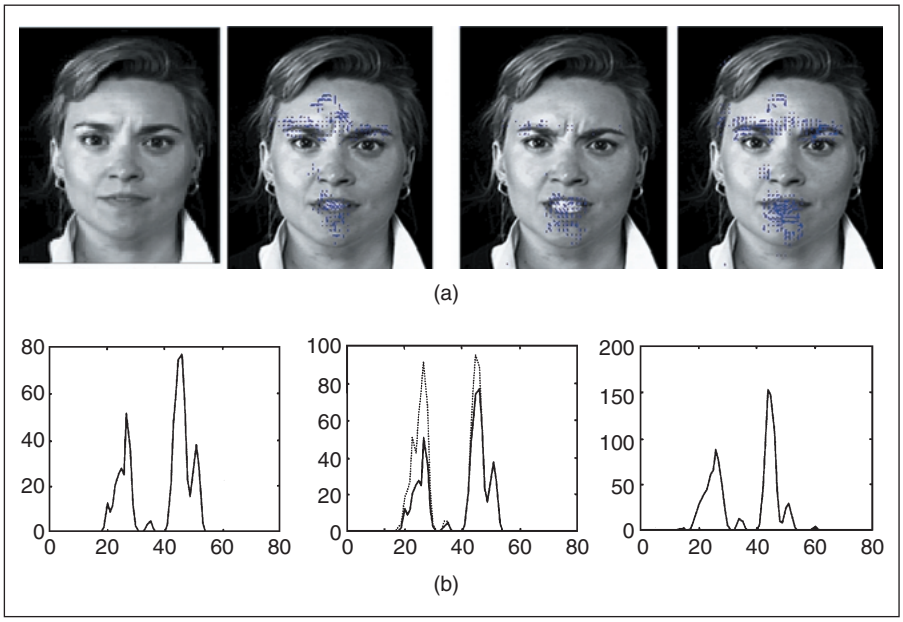
Simulation

We show that given prior normalization (as in the former section), and a preprocessing stage for face detection, relatively straightforward techniques can give reasonable classification. We estimate the optical flow directly from facial pixel values. The motion field is computed only in facial areas where substantial movement has occurred [114]: Standard learning algorithms use the resulting descriptions.

▲ Let F_k and F_{k+1} be the neutral and "apex" frames respectively, in which the face has already been detected by one of the techniques previously mentioned and normalized. Each pixel $p_k(x, y)$at the kth frame is described through its surrounding $2n \times 2n$block $b_k(x, y)$, and it is associated with the following error:



▲ *8. Discrete radon transform.*



▲ *9. Anger. (a) Frame 01, motion between frames 01 and 03, 03 and apex, apex and release; (b) radon transformation of motion amplitude at angle 90° between frames 01 and 03, 01 and apex (dotted line), apex and release.*

where t and u denote the x and y axes rotated by angle θ counterclockwise, as shown in Fig. 8.

The projections on two different angles, 0° and 90°, called "signatures" can be used as features for discriminating different expressions. To illustrate this fact, we used the (rather small) Yale database and obtained good classification performance, with respect to happy, surprised, sad, and sleepy expressions, ranging from 82% to 87.5%, using a correlation matching scheme and a neural network classifier respectively [114].

Fully Dynamic Techniques: Approaches to extracting facial emotions from image sequences fall into three classes which are described next. Of particular interest is the MPEG-4 framework, examined separately in the end.

$$e_k(x,y) = |b_k(x,y) - b_{k+1}(x,y)| = \sum_{l=-n}^{n} \sum_{m=-n}^{n} |p_k(x+l,y+m) - p_{k+1}(x+l,y+m)|. \quad (1)$$

▲ Motion vectors are calculated only for blocks with significant $e_k(x,y)$, using an appropriate image-dependent thresholding.

The motion vector $\hat{\mathbf{v}}_k(x,y)$ of block $b_k(x,y)$ is computed using block matching in a neighborhood of block $b_{k+1}(x,y)$ according to

$$\hat{\mathbf{v}}_k(x,y) = (\hat{v}_x, \hat{v}_y) = \underset{(v_x,v_y)\in Q}{\operatorname{argmin}} \sum_{l=-n}^{n} \sum_{m=-n}^{n} |p_k(x+l,y+m) - p_{k+1}(x+l-v_x, y+m-v_y)| \quad (2)$$

where $Q = \{-q,\ldots,q\} \times \{-q,\ldots,q\}$ is the search area. To decrease execution time, logarithmic search is employed, using a limited subset of combinations $(\hat{v}_x, \hat{v}_y) \in Q$ for searching. "Noisy" motion vectors, i.e., poorly estimated pixel motion, inevitably arise due to the simplicity of motion estimation; to account for this, median filtering is applied first to motion vector phases (in the form of directional filtering) and then to their magnitude.

The motion vector at position (x,y) can be expressed as $\hat{\mathbf{v}}_k(x,y) = a_k(x,y)e^{j\phi_k(x,y)}$. The discrete Radon transform of the motion vector magnitude, at angle θ, is then given by

$$R(\theta) = \sum_{u=-\infty}^{\infty} a_k(x,y)\Big|_{x=t\cos\theta - u\sin\theta,\, y=t\sin\theta + u\cos\theta} \quad (3)$$

Optical Flow-Based Approach

The optical flow based approach uses dense motion fields computed in selected areas of the face, such as the mouth and eyes; it tries to map these motion vectors to facial emotions using motion templates which have been extracted by summing over a set of test motion fields [114], [116]. A coarse-to-fine strategy using a wavelet motion model can be used for facial motion estimation, in cases where the displacement vectors between successive frames can become large [117]. A problem in these approaches, which in general are computationally intensive, is caused by the inherent noise of the local estimates of motion vectors, which may result in degradation of the recognition performance.

Ohya et al. [118]-[120] applied hidden Markov models (HMM) to extract feature vectors. Motivated by the ability of HMMs to deal with time sequences and to provide time scale invariance, as well as by their learning capabilities, they assigned the condition of facial muscles to a hidden state of the model for each expression. In [118] they used the wavelet transform to extract features from facial images. A sequence of feature vectors was obtained in different frequency bands of the image, by averaging the power of these bands in the areas corresponding to the eyes and the mouth. In [119] continuous output probabilities were assigned to the HMM observations and phase information was added to the feature vectors. In [120] feature vectors were obtained in two steps. Velocity vectors were estimated between every two successive frames using an optical flow estimation algorithm. Then a two-dimensional Fourier transform was applied to the velocity vector field at the regions around the eyes

and the mouth. The coefficients corresponding to low frequencies were selected to form the feature vectors.

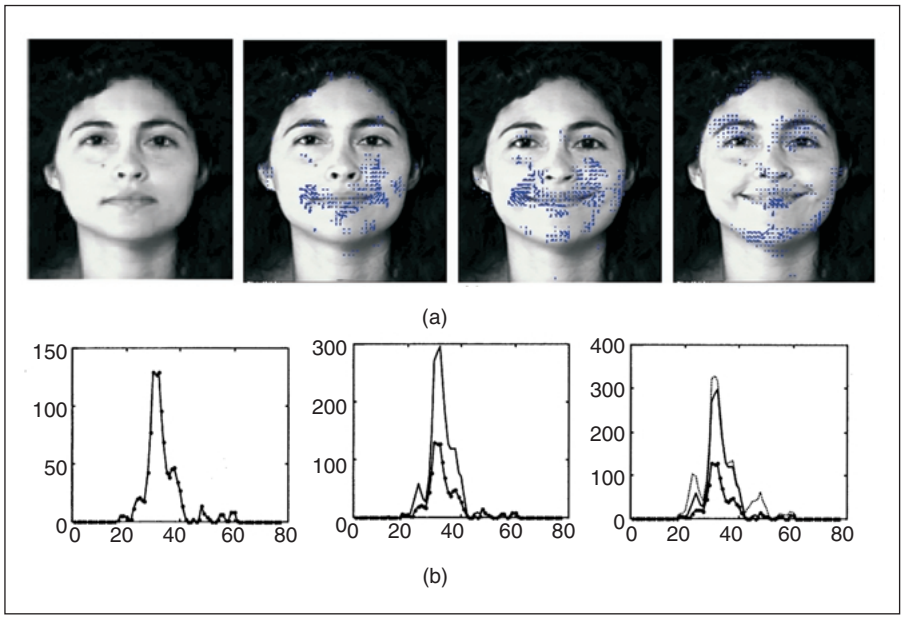
Vector quantization was used for symbolization in [118], enhanced by a category separation procedure. Experiments for recognizing the six archetypal expressions provided a recognition rate of 84.1%, which was slightly improved in [119]. A recognition rate of 93%, including successful identification of multiple expressions in sequential order, was obtained in [120].

Feature Tracking Approach

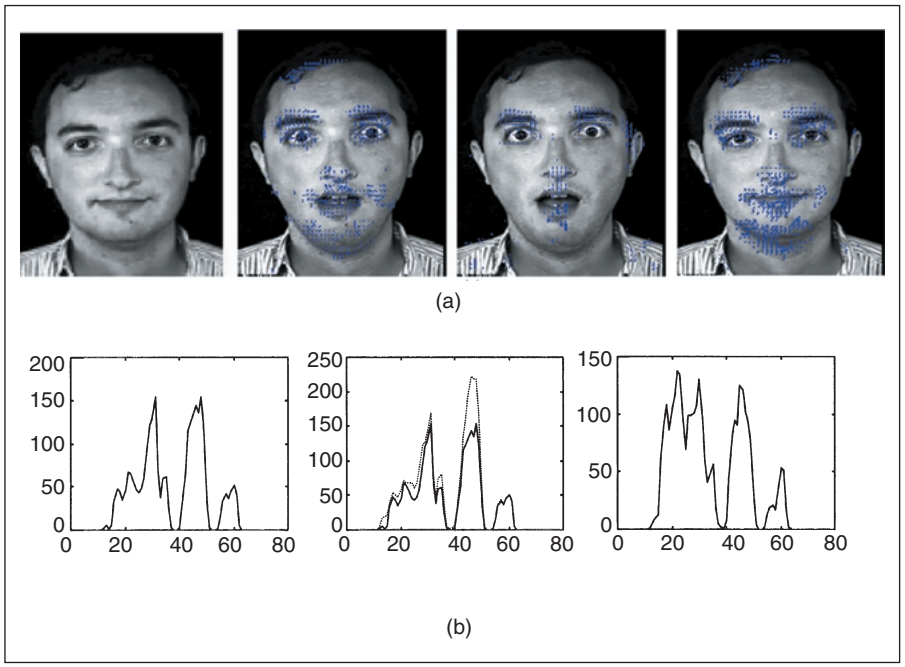
In the second approach, motion estimates are obtained only over a selected set of prominent features in the scene. Analysis is performed in two steps: first each image frame of a video sequence is processed to detect prominent features, such as edges, corner-like and high-level patterns like eyes, brows, nose, and mouth [88], [142]-[145], followed by analysis of the image motion [146]. In particular, the movement of features can be tracked between frames using Lucas-Kanade's optical flow algorithm [89], which has high tracking accuracy. The advantage of this approach is efficiency, due to the great reduction of image data prior to motion analysis; on the other hand, it is not certain that the feature points which can be extracted suffice for the emotion recognition task.

Yacoob [114], [121], extending his previous work, used dense sequences (up to 30 frames per second) to capture expressions over time. The focus was on near frontal facial views and on motion associated with the edges of mouth, eyes, and eyebrows. Optical flow was computed, with subpixel accuracy, only at points with high gradient in each frame [111]. On a sample of 46 image sequences of 32 subjects, the system achieved recognition rates of 86% for happiness, 94% for surprise, 92% for anger, 86% for fear, 80% for sadness, and 92% for disgust. Blinking detection was achieved at 65% of cases. Some confusion of expressions occurred between the pairs of fear and surprise, anger and disgust, sadness and surprise; these pairs are in proximity in space and share common action units, so that human judges make similar errors.

Using the same features as in [113], Rosenblum and Yacoob [122] proposed a radial basis function (RBF) network architecture to classify facial expressions. They used a hierarchical approach, which at the highest level identified emotions, at the mid-level determined motion of facial features, and at the lowest level recovered motion directions. Correct recognition was 76%. Related neural network based techniques were developed by Thalmann et al. [123].



▲ *10. Happiness. (a) Frame 01, motion between frames 01 and 05, 05 and 09, 09 and apex; (b) radon transformation of motion amplitude at angle 90° between frames 01 and 05 (dashed line), 01 and 09, 01 and apex (dotted line).*

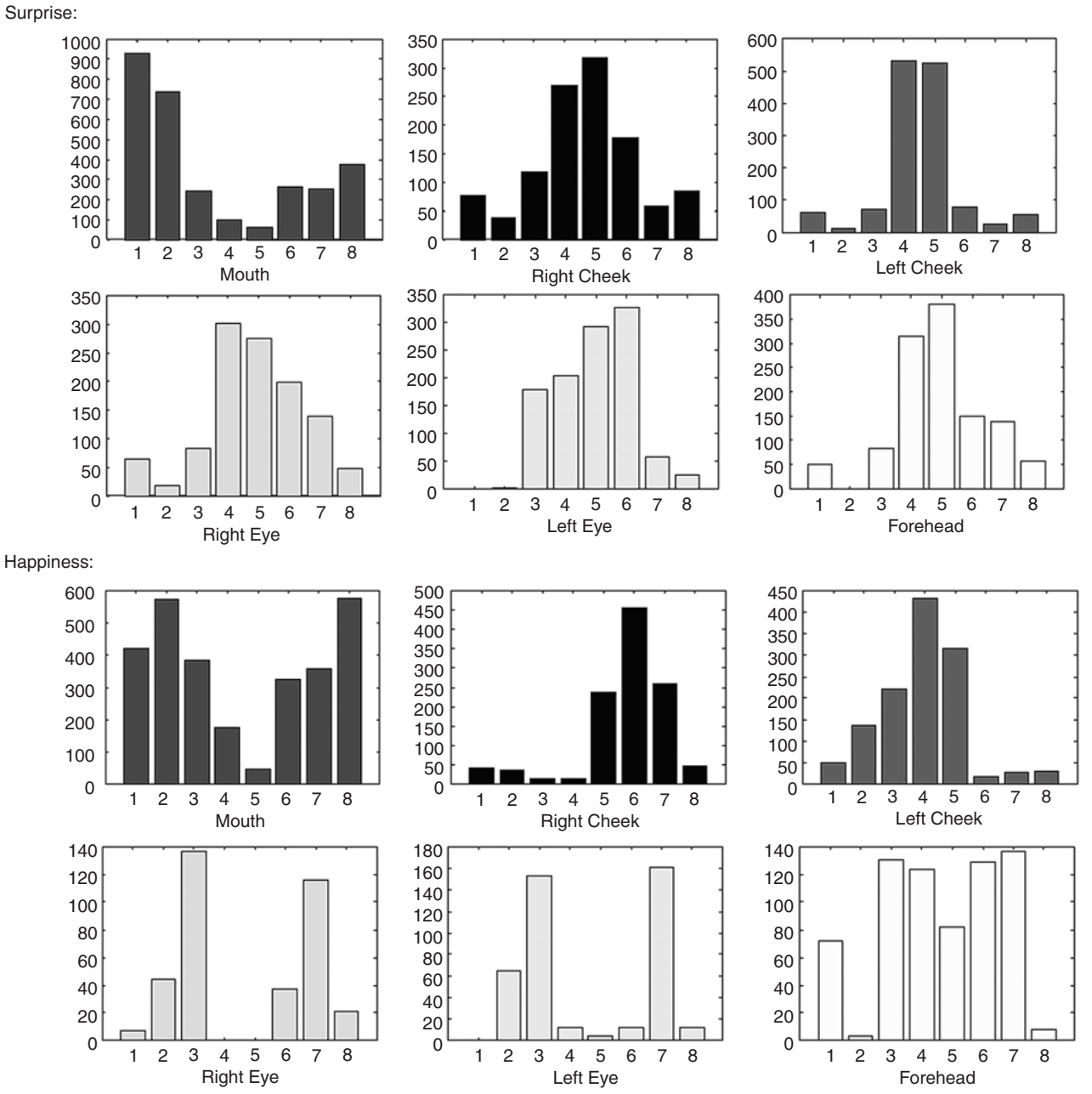


▲ *11. Surprise. (a) Frame 01, motion between frames 01 and 04, 04 and apex, apex and release; (b) radon transformation of motion amplitude at angle 90° between frames 01 and 04, 01 and apex (dotted line), apex and release.*

Model Alignment Approach

The third approach aligns a 3-D model of the face and head to the image data to estimate both object motion and orientation (pose). A series of publications by Essa and Pentland [124]-[126] dealt with tracking facial expressions over time, using an extended FACS representation and matching spatio-temporal motion-energy templates of the whole face to the motion pattern.

Based on the work of Platt and Badler [127], who had created a mesh based on isoparametric triangular shell elements and on that of Waters [99] who had proposed a muscle model in a dynamic framework, they created a dynamic model of the face, describing the elastic nature of facial skin and the anatomical nature of facial muscles. The transformations creating this model and its control parameters provided the necessary features to describe facial expressions through muscle actuation. Each expression was divided into three distinct phases, i.e., application, release, and relaxation. Feature extraction was similarly divided into three distinct time intervals and normalized by the temporal course of the expression. Using a small database of 52 sequences, a recognition rate of about 98% was obtained.



▲ *12. Energy of movement at the eight main directions of face areas. Directions: 1: Down, 2: Left Down, 3: Left, 4: Left Up, 5: Up, 6: Right Up, 7: Right, and 8: Right Down.*

Simulation

As before, we have integrated key ideas from the literature into a relatively straightforward system that learns expression classification based on optical flow. Flow is estimated taking into account consecutive frames; the computed motion field is accumulated over all time periods, as shown in Figs. 9(a), 10(a), and 11(a) for the expressions of anger, happiness, and surprise, respectively. Corresponding "signatures" are obtained by applying the Radon transform to the motion field in angles 0° and 90°; the latter case is shown in Figs. 9(b), 10(b), and 11(b). Motion energies, corresponding to movement in eight different directions of important facial parts, are also extracted, as shown in Fig. 12 for the expressions of happiness and surprise. "Signatures" and motion energies form a feature vector sequence, which can feed six HMMs for classification of the archetypal expressions [128]; HMMs perform the required temporal analysis, also accounting for time-scale variations.

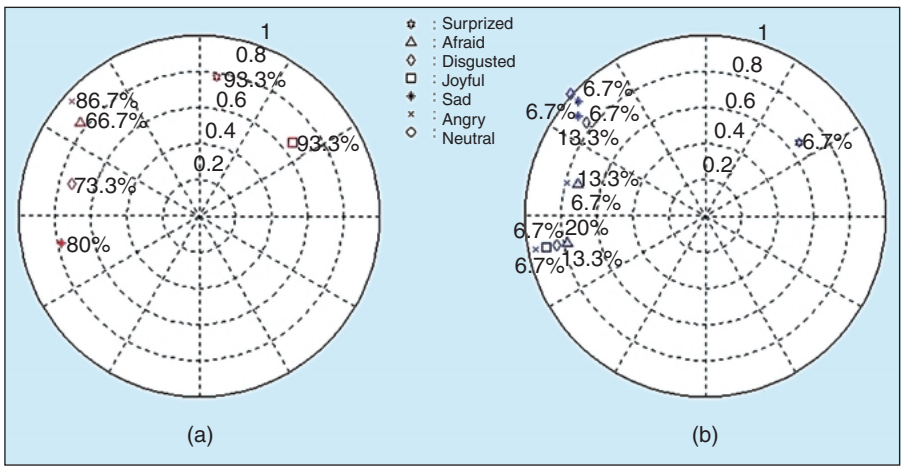
The results obtained by the HMMs were explored using the activation—evaluation space, described earlier. The key question was whether the variables learned by the HMMs relate systematically to underlying dimensions rather than to wholly discrete categories. Figure 13 suggests that they do to at least some extent. The majority of misclassifications (especially in the categories of anger, sadness, fear, disgust, and surprise) remain in the same quadrant as the correct response. Hence even misclassifications generally convey broad information about emotional state. Since that occurs even though training does not use information about activation/evaluation space, it suggests a natural compatibility between some dimensions of facial movement and dimensions of emotion. Future research will seek to understand that compatibility and capitalize on it.

Feature Tracking in the MPEG-4 Framework

The natural way to develop feature tracking techniques is to use the standard MPEG-4 FDP points. On the basis of previous work, we have identified a subset of the FDP 3-D feature point, shown in Fig. 5, that appears relevant to emotion. Based on that, we have created a relation between FAPs that can be used for the description of six archetypal facial expressions [148] and the selected FDP subset. Table 12 presents the features that represent this relation, which are the time derivatives of the distances between the selected FDP points, normalized by the corresponding FAPU. The fifteenth feature is not related to a distance itself, but tries to capture the gesture of vertical wrinkles created in the area between the inner parts of the eyebrows. The above features define the positive intensities for the FAP set as well as the development of the expressions towards their "apex" state.

We have examined the potential of that approach in two separate simulations.

In the first, we used sequences obtained from the MIT Media Lab, which show standard archetypal emotions—happiness, surprise, anger, and disgust. Faces were detected in the image frames using color and shape information, as mentioned earlier (see also [147]). The technique presented in [149] was then used to detect the relevant FDP subset. Accurate detection of the FDP points, however, was assisted by human intervention in many cases. Then, for each picture that illustrated a face in an emotional state, the feature vector corresponding to FAPs was computed. A neural network architecture was trained and then used to classify the feature vectors in one of the above categories in a frame by frame basis; we were unable to explore temporal classification of the whole sequences, due to the limited number of available sequences in the database. The results were projected onto the activation—evaluation space and are shown in Fig. 14(a). The results are in general good and similar to the ones obtained by the HMM approach. Using capabilities of neural networks [150], we were able to evaluate the contribution of each particular feature of the 15-tuble feature vector in the obtained results. Figure 14(b) indicates that about eight FAPs, related to the eyebrow and lip points, mainly contributed to the classification of the above expressions. The role of the wrinkle detection feature was found important for correct classification of the anger expression. The results also depict that features sensitive to accurate detection of the FDP points, such as *open_eyelid* and *close_eylid*, seem to be ignored by the neural classifier. Anomalously, only one of the components of symmetrical features is taken into account, i.e., the contribution of the *raise_l_i_eyebrow* is much higher than that of *raise_r_i_eyebrow*. These results suggest the possibility of coarse but robust description of emotions, using



▲ *13. (a) Correctly and (b) erroneously classified results using an HMM, projected to the activation-evaluation space.*

a relatively small number of FAPs and mapping results onto activation-evaluation space.

The second experiment considered the possibility of subtler discriminations, involving emotional states other than archetypal emotions. The emotional states considered were amusement, happiness, and excitement. Stimuli were drawn from two sources, the MIT facial database and a pilot database of selected extracts from BBC TV programs. Using the same procedure as before, we trained a neural network to classify the feature vectors in one of the three categories. Results are shown in Fig. 15(a), indicating that amusement and happiness were classified satisfactorily, whereas excitement was almost equally split between the three categories. That underlines the importance of looking beyond archetypal emotions. Feature sets developed to discriminate them may not support discriminations among other emotional states that are at least as common and important in everyday life.

Table 12. Description of FAP Set Using a Subset of the MPEG-4 FDP Set.

FAP Name	Features Used for the Description	Positive Intensity
Squeeze_l_eyebrow	$f_1 = \frac{s(1,3)}{ESo}, \frac{df_1}{dt}$	$\frac{df_1}{dt} < 0$
Squeeze_r_eyebrow	$f_2 = \frac{s(4,6)}{ESo}, \frac{df_2}{dt}$	$\frac{df_2}{dt} < 0$
raise_u_midlip	$f_3 = \frac{s(16,30)}{ENSo}, \frac{df_3}{dt}$	$\frac{df_3}{dt} < 0$
raise_l_midlip	$f_4 = \frac{s(16,33)}{ENSo}, \frac{df_4}{dt}$	$\frac{df_4}{dt} < 0$
raise_l_i_eyebrow	$f_5 = \frac{s(3,8)}{ENSo}, \frac{df_5}{dt}$	$\frac{df_5}{dt} > 0$
raise_r_i_eyebrow	$f_6 = \frac{s(6,12)}{ENSo}, \frac{df_6}{dt}$	$\frac{df_6}{dt} > 0$
raise_l_o_eyebrow	$f_7 = \frac{s(1,7)}{ENSo}, \frac{df_7}{dt}$	$\frac{df_7}{dt} > 0$
raise_r_o_eyebrow	$f_8 = \frac{s(4,11)}{ENSo}, \frac{df_8}{dt}$	$\frac{df_8}{dt} > 0$
raise_l_m_eyebrow	$f_9 = \frac{s(2,7)}{ENSo}, \frac{df_9}{dt}$	$\frac{df_9}{dt} > 0$
raise_r_m_eyebrow	$f_{10} = \frac{s(5,11)}{ENSo}, \frac{df_{10}}{dt}$	$\frac{df_{10}}{dt} > 0$
open_jaw	$f_{11} = \frac{s(16,33)}{ENSo}, \frac{df_{11}}{dt}$	$\frac{df_{11}}{dt} > 0$
close_upper_l_eyelid – close_lower_l_eyelid	$f_{12} = \frac{s(9,10)}{ENSo}, \frac{df_{12}}{dt}$	$\frac{df_{12}}{dt} < 0$
close_upper_r_eyelid – close_lower_r_eyelid	$f_{13} = \frac{s(13,14)}{ENSo}, \frac{df_{13}}{dt}$	$\frac{df_{13}}{dt} < 0$
stretch_l_cornerlip – stretch_r_cornerlip	$f_{14} = \frac{s(28,29)}{ESo}, \frac{df_{14}}{dt}$	$\frac{df_{14}}{dt} > 0$
Vertical_wrinkles between eyebrows	$f_{15} = s'(3,6), \frac{df_{15}}{dt}$	$\frac{df_{15}}{dt} > 0$

Note: $s(i,j)$ = Euclidean distance between FDP points i and j, $\{ESo, ENSo\}$ = Horizontal and vertical distances used for normalization, and $s'(3,6)$ is the maximum difference between pixel values along the line defined by the FDPs 3 and 6.

Natural Directions for Research

This section attempts to identify the natural directions for further research, identifying where existing approaches are incomplete and how integration among them could be accomplished.

Almost all of the work mentioned in "Computational Studies of Facial Expression Recognition" deals with archetypal emotions. It focuses on facial features and models that can discriminate among that limited range of expressions. The available data also reflect the emphasis attached to this type of classification. It is a natural priority to make use of a wider range of categories and richer output representations, as suggested by the activation-evaluation representation and the BEEVer emotional description. That interacts with investigation of potential midlevel representations and extracted features, because selection depends on the variety of situations to be described. The MPEG-4 framework is clearly relevant because it will have a crucial role in forthcoming HCI applications and environments. Automatic extraction of the FDP parameters should improve rapidly, and that will permit the generation of more powerful FAP representations. Because of the continuity of emotion space and the uncertainty involved in the feature estimation process, it is natural to apply fuzzy logic or neurofuzzy approaches. Similarly, hybrid approaches offer natural ways of coding and enriching prior knowledge, and adapting it to accommodate the different expressive styles of specific users. Moreover, developing systems which can combine and analyze both visual and speech input data is crucial. Since HMMs have been successfully applied to both types of input, they are a natural element of this combined framework.

All of those developments hinge on the construction of appropriate databases, preferably audio/visual. That topic is further discussed in the next section.

Training and Test Material

If emotion recognition systems use learning architectures, such as neural networks, then adequate training and test material are as fundamental as feature extraction. Training and test material need to contain two streams. One—the input stream—describes visual and acoustic in-

puts directly. The other—the output stream—describes episodes in the signal stream in terms related to emotion. The descriptions in the output stream will be called interpretations. This section examines the principles relevant to selecting the streams, resources that are currently available, and approaches to construct new material.

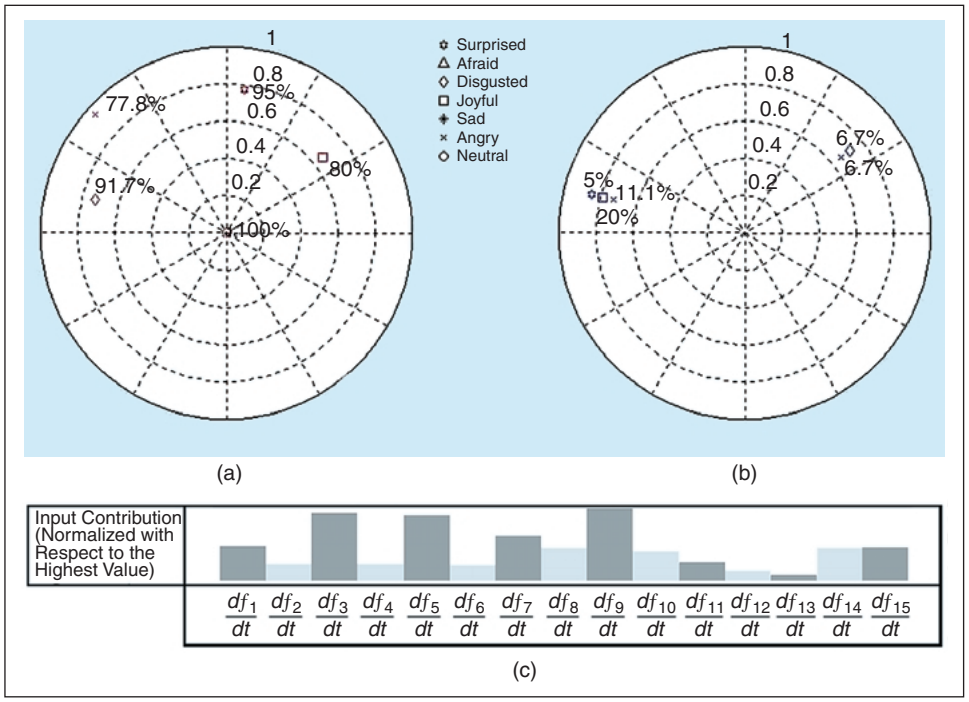
Existing Material

This subsection reviews existing collections of material with at least some connection to the expression of emotion in human faces or vocal behavior. The Web contains many sites that refer to research on topics related to emotion and facial and speech information, including a number of databases with at least some relevance to our goals. A full description of associated databases can be found in [151].

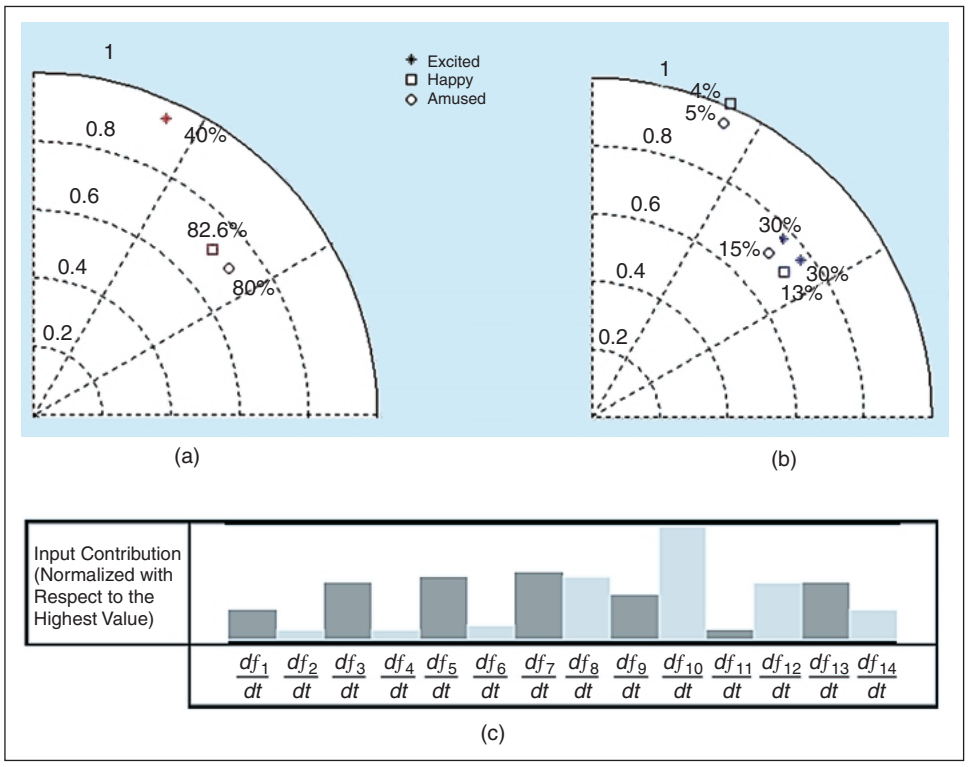
Sources Dealing with the Study of Human Emotion

Many links concerned with emotion refer to sites where research in Psychology is carried out. These sites (e.g., the UC Berkeley Psychophysiology Lab, the UC Santa Cruz Perceptual Science Lab, the Geneva Emotion Research Group) tend to deal with the theory of emotional expression and its relation to cognition. They give interesting details about experiments on emotional behavior and the detection of emotion, using speech and facial animation as inputs. They also provide rich bibliographic references. However, they tend not to make their signal sources publicly available.

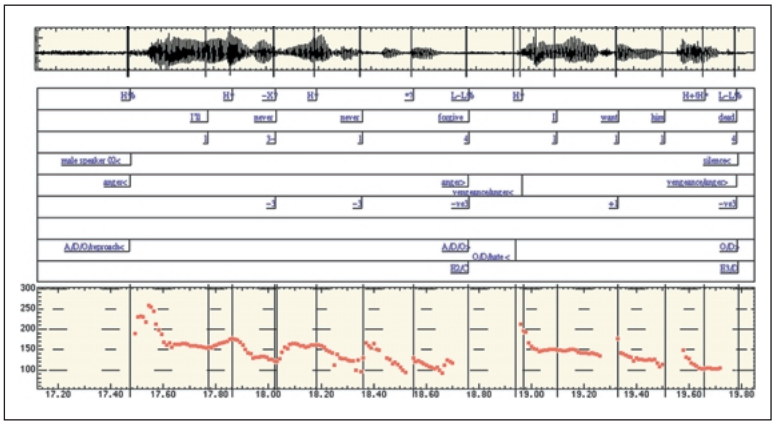
Four sites are particularly relevant. An interesting overview of models and approaches used historically in this area of human psychology, as well as a list of related links, is available at the site of the Salk Institute for Biological Studies (http://emotion.salk.edu/emotion.html). Recent techniques for facial expression analysis and acoustical profiles in vocal emotion are published by a team at Geneva (http:// www.unige.ch/fapse/emotion/welcome.html). The Berkeley site pro-



▲ *14. (a) Correctly and (b) erroneously categorized results obtained using the MPEG-4 framework, projected to the activation-evaluation space; (c) contribution of the particular features to the classification task (see Table 12 for the particular features).*



▲ *15. (a) Correctly and (b) erroneously categorised results concerning variations of the happy emotion, projected to the activation-evaluation space; (c) contribution of the particular features to the classification task (see Table 12 for the particular features).*



▲ *16. An annotated sample from the Reading/Leeds project. A number of interesting points can be made about the format.*

▲ *17. A sample of Yale face database.*

vides suggestions (http:// socrates.berkeley.edu/~ucbp1/research.html) about eliciting emotion in laboratory settings.

On the other hand, most sites deal with topics only indirectly related to emotion and its audiovisual expression. They reflect the enormous amount of interest that has been stimulated by research on man-machine interaction, computer vision, medical applications, lipreading, videoconference, face synthesis, and so on. Most of them refer to the theoretical basis of their approach or examine issues like relationships between facial expression, aging, and attractiveness, trying to describe features which could lead a machine to detect cues to these human characteristics (see the links proposed by the UCSC Perceptual Science Lab, http://mambo.ucsc.edu/). Emotional content of faces and voices is often an issue, but rarely the main target of their research. A few sites make their signals freely available as databases. Video sequences containing facial expressions can be downloaded from various web sites as detailed below whereas speech materials tend to be more scarcely represented.

In the summary that follows, we try to give as much information as possible about material that is related to our aims. More details about the material and its exact location can be found in [151].

Speech

Corpus linguistics has been a major research area in the past decade, and it has produced a number of substantial speech databases, in many languages. Several of them include emotion-related material.

An emotional speech database has been complied for Danish (DES, 1997) but no details are yet available. A CD-ROM contains 48kHz sampled audio files and a phonotypical SAMPA transcription of 30 minutes of speech [involving two words (yes and no), nine sentences (four questions), and two passages] performed by four actors believed able to convey a number of emotions (neutral, surprise, happiness, sadness, anger). It is available at the Center for Person Kommunikation of Aalborg. Another corpus partially oriented to the analysis of emotional speech is called GRONINGEN (ELRA corpus S0020), including over 20 hours of Dutch read speech material from 238 speakers in CD-ROM.

For English, a joint research project between The Speech Laboratory at the University of Reading and The Department of Psychology at the University of Leeds carried out the Emotion in Speech Project [51]. Samples are organized in a database. It is to be released on CD-ROM, but is not available yet.

Figure 16 shows an annotated sample from the Reading/Leeds project. A number of interesting points can be made about the format. The input stream is augmented with prosodic features described using the standard ToBI system [200]. Although the result is included in the database, they conclude that it is too phonologically oriented to permit detailed representation of the phonetic and paralinguistic information that is relevant to analysis of emotional speech. As a result they added descriptors based on the prosodic and paralinguistic feature system devised by Crystal [37]. The output stream also goes beyond the classification system that phoneticians use in describing attitudes and emotions transmitted by vocal means. Both points reinforce conclusions drawn in earlier sections.

Faces: Static Images

Faces have been a focus of research in several disciplines, and databases that present pictures of them are offered by many labs all over the world. These collections could represent a source of test material, but are not always emotionally expressive, or associated with suitable descriptors of emotional state.

Electronic Collections

There are many collections of static pictures that show faces under systematically varied conditions of illumination, scale, and head orientation, but very few consider emotional variables systematically. There are examples, which do portray emotion, but bringing together samples from various databases with nonuniform format is not an ideal procedure in terms of practicality or consistency. Databases containing emotion-related material that are freely and immediately available include the following.

The one, provided by Yale (http://cvc.yale.edu/projects/yalefaces/ yalefaces.html), includes 165 gray scale images (size 6.4 MB) in GIF format of 15 individuals. There are 11 images per subject, one per different facial expression or configuration: center-light, with glasses, happy, left-light, without glasses, normal, right-light, sad, sleepy, surprised, and wink. The database is publicly available for noncommercial use. Figure 17 shows a sample of a Yale sequence. It illustrates a recurring problem in the area. A suitable term for the images might be "staged." Expressions like these might be encountered in politics, storytelling, or the theater, but intuition suggests that it would be very disturbing to encounter them during something that we regarded as a sincere interaction at home or at work.

Figure 18 shows a sample of expressions made available by the Geneva group. The contrast with Fig. 17 is very striking: it is obvious that the pictures show people genuinely engaged in emotional behavior. However, the available data set is small. The group reports having records of elicited expressions, but there is no explicit intention to make them widely available.

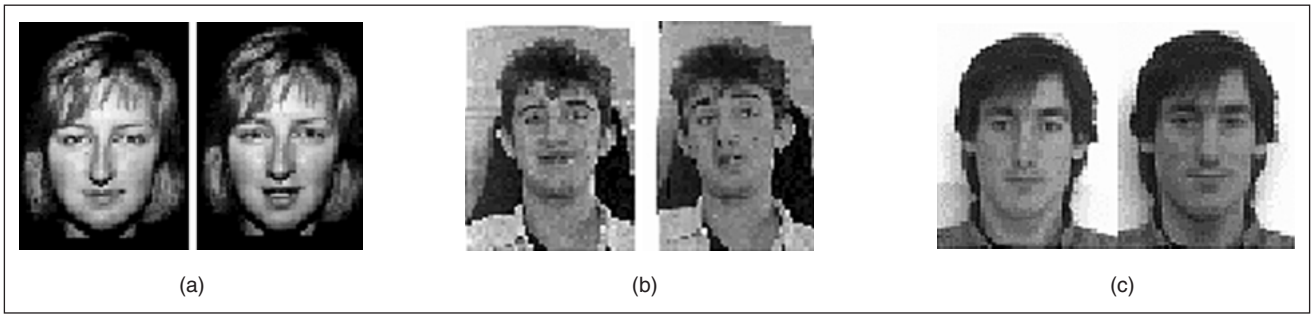
The ORL Database of Faces (ftp://ftp.orl.co.uk:pub/data/orl_faces.zip) contains a set of face images taken between April 1992 and April 1994 at ORL. The database was used in the context of a face recognition project carried out in collaboration with the Speech, Vision and Robotics Group of the Cambridge University Engineering Department. There are ten different images of each of 40 distinct subjects. The images vary the lighting, facial details (glasses /no glasses), and aspects of facial expression which are at least broadly relevant to emotion—open /closed eyes, smiling /not smiling. All of the images were taken against a dark homogeneous background with the subjects in an upright, frontal position (with tolerance for some side movement). The files are in PGM format, each image containing 92×112 pixels, with 256 grey levels per pixel.

Another important collection of face images is available at the site of the PICS database at the University of Stirling (http://pics.psych.stir.ac.uk/cgi-bin/PICS/pics.cgi), within a larger collection of various other images. Samples of face images are available in demo version. All the collections can be downloaded after registration as tar-compressed files. Among the most promising are those of a first database of 313 images, where faces show three different expressions each [see Fig. 19(a)] and those of a second da-

(a)

(b)

▲ *18. (a) Examples from the Geneva project. (b) A sample of expressions made available by the Geneva group.*



▲ *19. Samples of images from PICS database*

tabase of 493 images and two expressions per subject [see Fig. 19(b)]. A third database is composed by 689 face images [see Fig. 19(c)] with four expressions represented.

Printed Collections

The classic collection of photographs showing facial emotion was published by Ekman and Friesen [152]. It can also be bought in electronic form. It is the natural reference source for computational research on static visual indicators of emotion. It is important, though, to recognize that the accompanying text makes it explicit that the pictures are anything but representative. They were posed, and stringently selected from a larger set, aiming at being consistent with Ekman and Friesen's theory. They are to everyday emotions rather as publicity photographs are to resorts in a variable climate.

The PSL site lists various other historically significant collections of face images or illustrated studies of emotion expression, e.g., Bulwer (1648), *The Deafe*; Duchenne de Boulogne (1862), *The Mechanism of Human Facial Expression*; Darwin (1872), *The Expression of the Emotions in Man and Animals*; Ermiane (1949), *Jeux Musculaires et Expressions du Visage*; and Ermiane and Gergerian (1978), *Album des Expressions du Visage*.

Faces: Videos

Relatively few sources contain samples of faces moving, and kinetic sequences which are emotionally characterized are even less common. Those which are available as freeware rarely exceed demonstrations with sequences of three or four frames.

The material that tends to be available at these sites consists of images produced by research software—e.g., for lip tracking or facial animation—instead of the original video sequences or images used for the analysis and/or the training. An impressive list of projects, carried out on these fields, are given at the PSL site (http:// mambo.ucsc.edu/psl/fanl.html). Samples of movies of faces expressing emotion are at the location (http://www.cs.cmu.edu/~face/) where they are used to describe a few examples of the application of the optical flow technique for six archetypal emotions (surprise, joy, anger, disgust, fear, sadness), while an interesting collection is in (ftp://whitechapel.media.mit.edu/pub/) describing smile, anger, disgust and surprise expressions.

Speech and Video

Material which combines speech and video is still rare. The M2VTS group, provided one of two speech-plus-video databases we have traced. Neither of them seems to be emotionally characterized.

M2VTS Multimodal Face Database (http://www.tele.ucl.ac.be/M2VTS) is a substantial database combining face and voice features, contained on three high density exabyte tapes (5 Gbyte per tape), focusing on research in multimodal biometric person authentication. It includes 37 different faces and provides five shots for each person taken at one week intervals or when drastic face changes occurred in the meantime. During each shot, people have been asked to count from zero to nine in their native language (most of the people are French speaking). The final format for the database is for images: 286×350 resolution, 25 Hz frame frequency/progressive format, 4:2:2 color components. Sound files (.raw) are encoded using raw data (no header). The format is 16 bit unsigned linear and the sampling frequency is 48 kHz. An Extended M2VTS Face Database also exists, including more than 1,000 GBytes of digital video sequences (http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb/). It is not clear which are the languages represented. Its short description only detailed that speakers were asked to read three English sentences which were written on a board positioned just below the camera. The subjects were asked to read at their normal pace, pause briefly at the end of each sentence, and read through the three sentences twice. The three sentences audio files, a total of 7080 files, are available on four CD-ROMS. The audio is stored in mono, 16bit, 32 kHz, PCM wave files.

Tulips 1.0 (ftp://ergo.ucsd.edu/pub/) is a small audio-visual database of 12 subjects saying the first four digits in English. Subjects are undergraduate students from the Cognitive Science Program at UCSD. The database contains both raw acoustic signal traces and cepstral processed files in PGM format, 30 frame/s. Audio files are in .au format. Video files are in PGM format, 100×75 pixel 8 bit gray level.

An archive called MIT-faces is also available at the site ftp://ergo.ucsd.edu/pub/MIT-faces/MIT-faces.mat. It contains a matrix with 48 rows and 36,000 columns. Each row is a different image. The columns go for 60×60 pixels per image. There are images of 16 subjects each of

which is displayed in three different illumination conditions for a total of 48 images.

Overview of Available Material

It is clear from this review that lack of suitable material is a major obstacle to progress. Where material is available, it needs to be scrutinized very carefully for ecological validity. That is why we now consider the issues involved in selecting and constructing training material.

Input Stream Construction

Recording

Recording suitable input material is not trivial. The speech community has recognized for many years that obtaining genuinely realistic recordings of any kind is a difficult problem, and psychologists trying to evoke genuine fear and anger have gone to lengths that no modern ethical code would allow [199]. This section outlines the main methods to elicit material, noting their advantages and disadvantages in relation to both constructing new material and evaluating preexisting sources.

Context-Free Simulations: This seems a suitable term for attempts to generate emotional behavior in a vacuum, such as posed photographs and emotionally ambiguous sentences spoken to convey a specified emotion. Context-free simulations are easy to carry out, and they can provide material which is controlled and balanced. Unfortunately, it is emphatically not natural. It is also likely to be biased with respect to structure. It is much easier to pose for a snapshot or speak a short sentence in a simulated emotion than it is to produce a sustained audiovisual simulation.

Reading: Reading material with appropriate emotional content is a step less artificial, mainly because well chosen passages can induce genuine emotion. For instance, passages used in [54] succeeded to the extent that several readers were unable or unwilling to read the passages conveying sadness and fear. However, the reading task incorporates severe constraints. Verbal content and phrasing reflect the intuitions of the person who wrote the passage, and facial expression is directly constrained by the need to keep the eyes on the text and to keep speaking (which constrains gestures with the mouth).

Prompting: A step less constrained again is providing a strongly emotive prompt. Prompts may be highly charged stories, extracts from film or television, or pieces of music. Various techniques using mental imagery have also been used to induce target emotional states. So far as speech is concerned, prompt techniques share some constraints with reading. People generally need an invitation to talk (usually about the prompt), and that tends to produce a set piece monologue. Prompts are also better at inducing some emotional states (sadness, anger, disgust, amusement) than others (love, serenity).

Games: Computer games have been increasingly used to induce emotion, particularly by the Geneva group. An environment called the Geneva Appraisal Manipulation allows them to generate experimental computer games that elicit emotion, with automatic data recording and questionnaires. While playing the experimental game, subjects are videotaped. Facial actions are then categorized in terms of FACS [98] and can be automatically matched to the corresponding game data, using a time code as a reference for both kinds of data [see Fig. 18(a)]. The approach does seem to elicit genuine emotion in a reasonably controlled way. However, it has various limitations. A minor one [which is noticeable in Fig. 18(b)] is that it constrains attention in a very distinctive way—subjects' gaze is generally focused on the screen. The major limitation is that at present, the technique elicits only facial expressions of emotion. That will change as voice input techniques for computers improve.

The expression of emotion may be quite context dependent, taking different forms in different settings.

Broadcasts: The Reading/Leeds project identified a large ready-made source of emotional speech, in the form of unscripted discussions on radio. For the audiovisual case, chat shows provide a comparable source. It would be naive to think that interactions in studio discussions were totally unaffected by the setting. Nevertheless, that kind of source seems more likely than the previous approaches to provide some expressions of quite strong emotions which are spontaneous and audiovisual.

Dialogue: Interactive situations are an important source for two reasons. First, emotion has a strong communicative function, and interactive contexts tend to encourage its expression. Second, ecologically valid accounts of emotion in speech need to consider dialogue, because it is the context in which speech usually occurs. Small groups of people who know each other well enough can talk freely or been set to talk on a subject that is likely to evoke emotional reactions. Results are considered later in this section.

These outlines make two points. First, they highlight sources of artificiality. Second, they highlight the intuition that the expression of emotion may be quite context dependent, taking different forms in different settings. As a result, both general and application-specific issues probably have to be considered in the selection of training material.

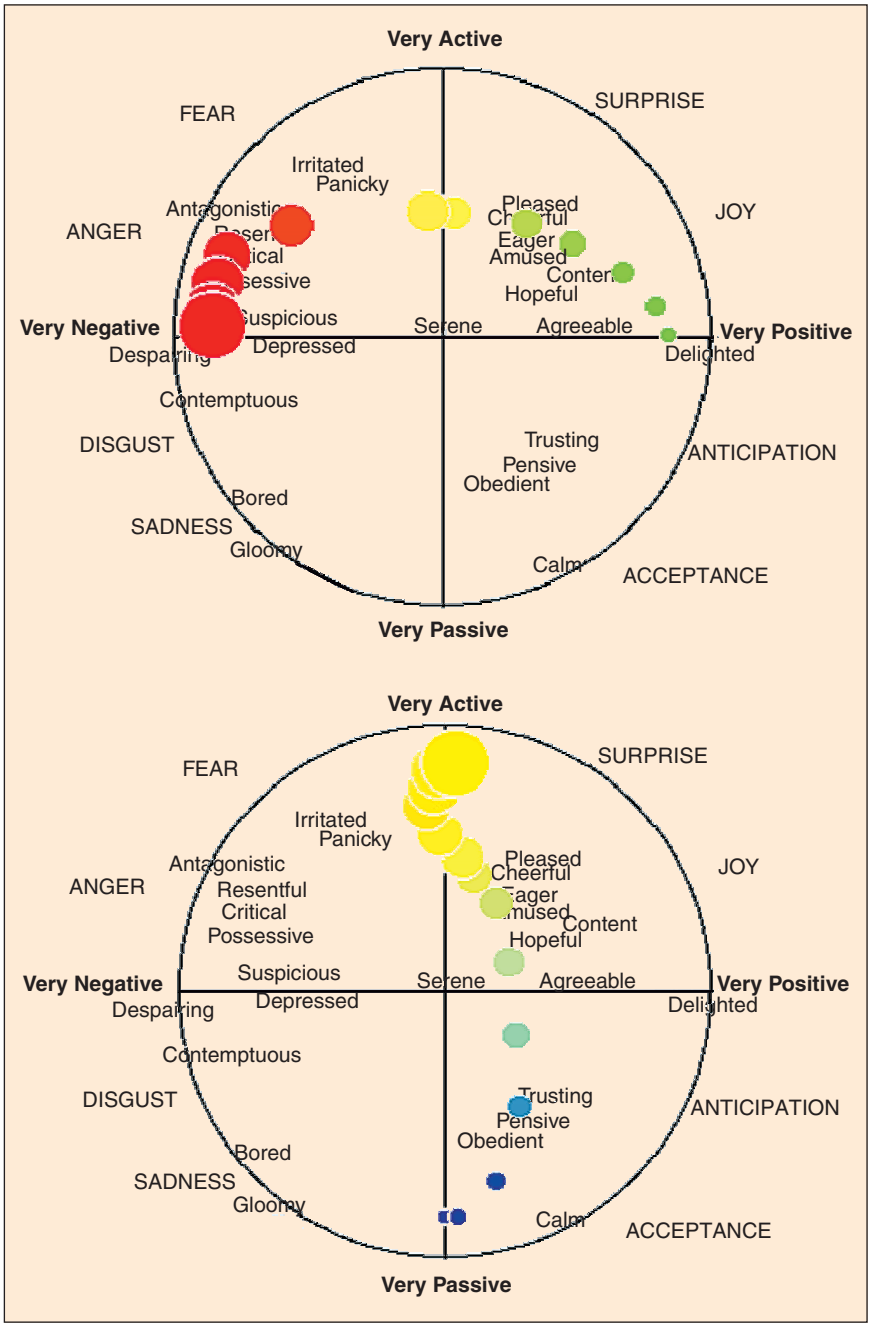
Output Stream Construction

An output stream consists of descriptions of emotional states. "A Descriptive Framework for Emotional States" reviewed the principles on which descriptions might be based. This section considers how they can be put into practice. It starts by considering requirements for an adequate database and then describes work aiming at satisfying them.

"Cause" and "Effect" Interpretations

Database structure is profoundly affected by the relative weights attached to cause-type interpretations, which purport to describe the state that gave rise to signs of emotion; and effect-type interpretations, which purport to describe the reaction that signs elicit in observers. These are closely related, because people are good at judging emotion. However, they diverge in some cases, such as deception or acting.

The psychological tradition is often strongly oriented towards cause-type interpretations, inducing particular, known emotions. That is appropriate in some contexts. However, in the present context, it is more appropriate to rely on effect-type interpretations—i.e., to describe the way human observers interpret the relevant signs of emotion, and hence to train artificial systems to judge emotion as they do. One reason is that requiring cause-type interpretations massively restricts the kinds of input sequence that can be used: for instance, it rules out material such as broadcasts where there is no access to the people who produced the emotion. A second reason is that even with access, cause-type interpretations are difficult to validate convincingly. For example, the fact that a performance was produced by asking someone to portray a particular emotion does not make the result a valid sample of that emotion. Effect-type interpretations, on the other hand, are straightforward to validate, by checking that observers assign them consistently. It makes sense to treat them as the primary source to be considered.



▲ *20. Examples of the display that a subject using Feeltrace sees at a particular instant during a response sequence—one chosen to show the negative/positive color coding, and the other to show the active/passive color coding.*

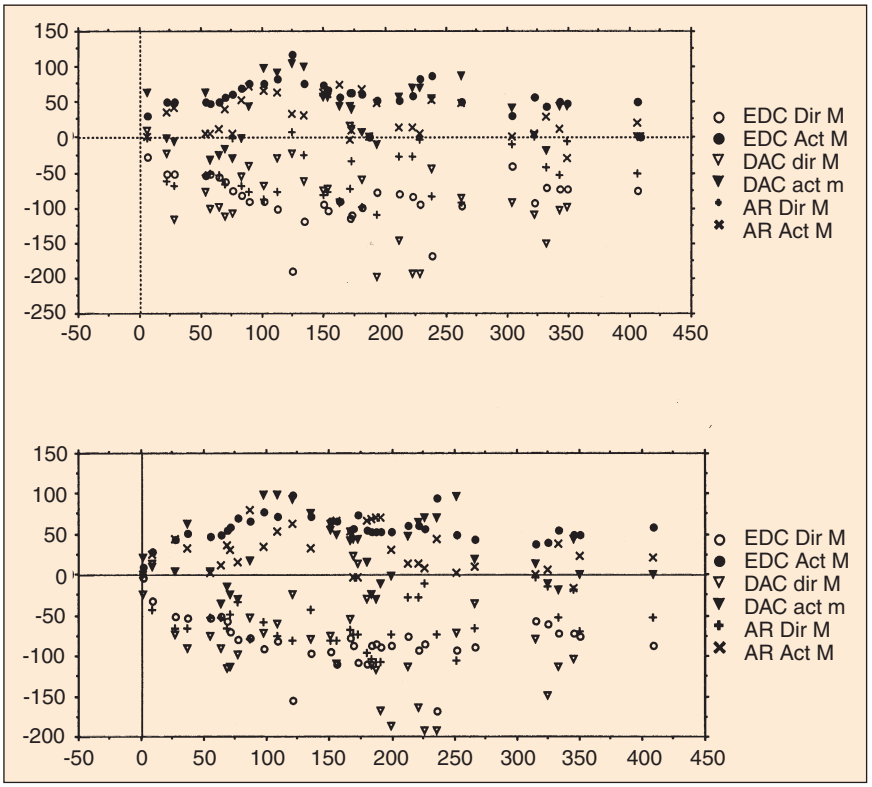
A closely linked issue is whether material should be used at all. Clearly, ineptly acted sequences should be excluded (unless the intention is to create a system that recognizes them as anomalous). Again, two types of questions are relevant—whether the sequence truly reflected emotions in the person who produced it and whether observers regard it as convincing.

Existing sources vary greatly in the level of validation that they describe. The Geneva group describes very full cause-type validation: situations are theoretically calculated to evoke particular emotions, and subjects rate their own emotional states. That makes the invaluable point that subjects who show moderate surface reactions may actually be more strongly affected subjectively than subjects who show strong surface reactions. The Geneva group have also carried out effect-type validation of vocal behavior and examined the features associated with convincing and unconvincing portrayals of emotion. At the other extreme, some sources make no mention of validation. Again, it will be assumed that the effect-type criterion has priority, i.e., a portrayal should be regarded as a legitimate source if it thoroughly convinces people. Clearly, though, independent information about the producer's emotional state is valuable if it is available.

Temporal Organization

Descriptions of emotional states need to be associated with units of some kind in the input stream. Finding suitable units is a nontrivial problem, reflecting questions that were raised earlier.

At one extreme, some critical kinds of evidence seem likely to come from brief episodes which offer strong signs of emotion; for instance, facial gestures lasting a few seconds. Some vocal signs of expression may be concentrated in brief episodes where a distinctive manner of speech comes through. At an intermediate level, statistical measures of speech need to be collected over a moderate period. It is not clear how long it needs to be to produce reliable information. Experience in related areas suggests that episodes of about ten seconds—which are roughly sentence-like—can be differentiated statistically. Other judgments seem to depend on considering larger units again. It seems unlikely that an emotion such as relief is identified without previous evidence of concern or worry, and triumph would be easier to identify given evidence that there had been conflict with another party who was now downcast. At the other extreme, judgments about moods or traits seem to depend on quite protracted samples.



▲ *21. Feeltrace results in pilot study.*

Training material needs to direct systems to recognize the time scale over which it is appropriate to form particular kinds of judgment and, in some cases at least, how relevant boundaries are marked.

Uniqueness and Uncertainty

It is tempting to assume that "good" data consists of episodes that have a definite emotional character. However, uncertainty is a salient feature of emotional life. Phrases such as "I don't know whether to laugh or cry" indicate that the person experiencing the emotion may be ambivalent. Shows like "Blind Date" regularly dramatize the fact that people find it difficult to recognize emotions related to attraction or dislike in another person.

Ideally, a database needs to be capable of reflecting at least the second kind of uncertainty. There are two ways of doing that. One is to incorporate measures of confidence within a single stream of data. The other is to attach alternative interpretations to episodes that elicit more than one type of reaction.

Eliciting Descriptions

Following the principles outlined above, we have developed a system for attaching effect-type descriptions of emotional state to an input stream. It has several elements.

"Feeltrace" is a computer system designed to let raters describe inputs in terms of activation-evaluation space. They are presented with a stimulus and specify its perceived emotional content by placing a mouse-controlled pointer within a circle that depicts of activation-evaluation space. Critically, moving the pointer allows them to track emotional content as it changes over time. The output is a list of coordinate pairs, one for activation and one for evaluation, each with a time stamp. The interval between pairs is of the order of 0.1 s.

Feeltrace incorporates multiple ways of conveying to the rater what a pointer position means. The main axes are marked and described as activation and evaluation. The color of the pointer is keyed to its position using a color code which people find reasonably intuitive [23]. The cursor is green in positions corresponding to highly positive emotional states, red in positions corresponding to highly negative emotional states yellow in positions corresponding to highly active emotional states, and blue in positions corresponding to very inactive emotional states. Around the edge of the circle are terms describing the archetypal emotion associated with that region of the space. Within the circle, selected words from the BEEVer list are printed at the point where their reported co-ordinates indicate that they belong. The dimension of time is represented indirectly, by keeping the circles associated with recent mouse positions on screen, but having them shrink gradually (as if the pointer left a trail of diminishing circles behind it). A substantial training session is given before raters use it. Figure 20 shows examples

of the display that a rater using Feeltrace sees at a particular instant during a response sequence—one chosen to show the negative/positive color coding, the other to show the active/passive color coding.

Feeltrace provides relatively fine resolution in time paired with coarse resolution in emotional space. To complement that, raters attach category labels from the BEEVer list to episodes with predefined boundaries. They must identify the best fitting label of 16 most frequently selected, and have the option of adding others from the remaining 24 if that adds clarification. Cause-type descriptions are also obtained when available. They are asked to rate their own emotions and to identify periods that they are genuinely caught up in the activity rather than "performing" for the recording.

Capturing Emotional States in Dialogue: A Case Study

We have reported preliminary work on the kind of material envisaged above [183]. Following the "social facilitation" approach, we arranged for groups of friends to choose topics that they felt strongly about and to discuss them for about an hour in a TV studio. The topics were religion, euthanasia, and communism. In each group, two people at a time were filmed. The third (unfilmed) person was asked to mediate and encourage rather than take a full part. After the sessions participants were asked to assess their own level of involvement. Most of them reported that they became caught up in the discussion for substantial periods and that at times they were emotionally very heated.

A selected passage was examined in depth. Three raters carried out Feeltrace ratings of a seven minute extract, presented auditorily. Its emotional tone was predominantly within a single Feeltrace quadrant (negative and active). The unit of analysis was a turn (i.e., a passage where one subject spoke without interruption from others). Figure 21 shows average ratings for each turn, separating out the two main speakers. It is reasonably easy to see which ratings are for activation and which are for evaluation because the former are almost all positive and the latter are almost all negative. Ratings show broad consistency, but there are also divergences. These divergences prompted the emphasis on training raters which was reported above, and the new procedure does appear to reduce them.

Phonetic analysis was carried out for 25 turns which provided contrasting, but reasonably consistent ratings. Prosodic features for each turn were measured using ASSESS. Feeltrace ratings were also summarized on a turnwise basis, using mean and standard deviation for each turn, on each dimension, and for each rater. There were correlations between ASSESS and Feeltrace measures, confirming that prosodic features do signal the emotional tone of an argument. However, the correlation patterns were not straightforward, and they were not the same for the two speakers. The strongest correlations involved change in emotional tone, suggesting that prosodic features signaled turns where a speaker was shifting ground emotionally. Change in activation level (measured by the standard deviation) was marked by low numbers of pitch movements per second and low variation in the amplitude of key segments. Change in evaluation (as captured by the standard deviation) was linked to variable pause length in both speakers and to a number of other features of pausing in one of the speakers.

Several features seemed to have variable significance. For example, one speaker signaled relatively positive evaluation with high standard deviation for the amplitude of falls and low minimum F0. In the other speaker, the same variables signaled change in evaluation. Similarly, one speaker signaled changing activation by raising the lower end of the pitch distribution, the other by raising the upper end. Initial tests related to visual signs were reported earlier. We are currently analyzing a larger variety of emotional states; the task is not generally straightforward. For example, the speaker who reported most emotion in the episodes studied for speech almost always retained a smile. Visual signs of the underlying emotion appear to reside in the manner of smile—from reasonably relaxed to forced.

That kind of material epitomizes the kind of challenge involved in recognizing emotion in the real world. People do give signs of their emotional state, and other people are able to detect them. But the signs are modulated, contextual, and multimodal. Building systems that can detect them depends on recordings that capture the way they operate, descriptions that express what they mean, and learning rules that can profit from that information; that is much more interesting than labeling posed photos.

Summary and Conclusions

Following the above, it can be concluded that developing artificial emotion detection systems ideally involves co-ordinated treatment of the following issues.

Signal Analysis for Speech

There is prima facie evidence that a wide range of speech features, mostly paralinguistic, have emotional significance. However, work is still needed on techniques for extracting these features. Techniques based on neural nets have been extensively used at this level and could be used more to set parameters within classical algorithms. There would probably be gains if the extraction process could exploit relevant linguistic information,phonetic or syntactic.

Signal Analysis for Faces

There is prima facie evidence that a range of facial gestures have emotional significance. The target-based approaches which are best known in psychology do not transfer easily to machine vision in real applications. Facial-gesture tracking approaches have produced promis-

ing results, but their psychological basis needs further exploration; they must also be tested on a large scale.

Effective Representations for Emotion

Describing emotion in an exclusive sense (i.e., cases of "pure" emotion) is very different from describing emotion in an inclusive sense (i.e., emotionality as a pervasive feature of life); and conceptions suggested by the first task do not transfer easily to the second. A range of techniques are potentially relevant to representing emotion in an inclusive sense, including continuous dimensions and schema-like logical structures. Ideally a representation of emotion should not be purely descriptive: it should also concern itself with predicting and/or prescribing actions. It should be also capable of modification through experience, as developmental and cross-cultural evidence indicate human representations are.

Appropriate Intervening Variables

Human judgments of emotion may proceed via intervening variables—referring to features of speech, facial gestures, and/or speaker state—rather than proceeding directly from signals. Describing intervening variables in symbolic terms will help to explain and reason about emotion-related judgments. We are currently working in this direction. Allowing suitable intervening variables to emerge through experience is certainly a challenge for computational theories of learning [153].

Acquiring Emotion-Related Information from Other Sources

Contemporary word recognition techniques support detection of words which have strong emotional loadings in continuous speech. Information from behavior and physical context are certainly relevant to emotional appraisal and can be obtained in at least some contexts. Active acquisition of information about emotionality is a clear possibility to be considered—e.g., asking "are you bored with this task?"

Integrating Evidence

Numerical methods of integrating evidence can generate good identification rates under some circumstances. In other circumstances it seems necessary to invoke logical techniques which examine possible explanations for observed effects, and discount them as evidence for X if explanation Y is known to apply—i.e., inferences are causal, adductive, and cancellable [154].

Emotion-Oriented World Representations

Cognitive theories highlight the connection between attributing an emotion and assessing how a person perceives the world in emotionally significant terms—as an assembly of obstacles, threats, boring, attractive, etc. Developing schemes which represent the world in emotion-oriented terms is a significant long term task which may lend itself to subsymbolic techniques. The task may be related to the well known that the meanings of everyday terms have an affective dimension [155].

Material

Some relevant databases already exist, but they have significant limitations. Hence developing suitable collections of emotional material is a priority. Material should be audiovisual and natural rather than acted. It should represent a wide range of emotional behavior, not simply archetypal emotions. It should cover a wide range of speakers and preferably cultures. Recordings should be accompanied by reference assessments of their emotional content. Traditional linguistic labelling is less critical. The goal is to obtain "live" material by *developing* scenarios which tend to elicit emotional behavior and which allow assessments of speaker emotions to govern actions. We are currently using the techniques presented in "Training and Test Material" to further develop this framework.

Acknowledgment

This work has been supported by the Training Mobility and Research project "PHYSTA: Principled Hybrid systems: Theory and Applications," 1998-2001, contract FMRX-CT97-0098 (EC DG 12).

Roddy Cowie received his B.A. in philosophy and psychology from Stirling University in 1972 and his D.Phil. from Sussex University in 1982. He became a Lecturer in psychology at Queen's, Belfast, in 1975 and Senior Lecturer in 1991. His research deals with relationships between human experience and computational models of perception. His publications include two monographs, six edited collections, and over 75 articles and chapters. He has organized conferences on machine and biological vision, deafness and University education, and most recently speech and emotion. He has also developed programs for identifying acoustic variables that correlate with human impressions of a speaker's personal attributes.

Ellen Douglas-Cowie received her B.A. in english studies from the New University of Ulster in 1972. She received her D.Phil from Ulster in 1980. She became a Lecturer in linguistics (based in the School of English) at Queen's University of Belfast, Senior Lecturer in 1991 and Head of School (1992 to present). Her research studies the characteristics that distinguish varieties of speech—clinical, social and stylistic—and includes seminal papers on sociolinguistics and deafened speech. She co-organized the recent International Speech Communication Association workshop on speech and emotion, and is currently developing a substantial database of emotional speech for the EC funded PHYSTA project on the recognition of emotion.

Nicolas Tsapatsoulis graduated from the Department of Electrical and Computer Engineering, the National Technical University of Athens in 1994, where he is currently working toward his Ph.D. degree. His current research interests lie in the areas of machine vision, image and video processing, neural networks and biomedical engineering. He is a member of the Technical Champers of Greece and Cyprus and a student member of IEEE Signal Processing and Computer societies. He has published six papers in international journals and 23 in proceedings of international conferences. Since 1995 he has participated in seven research projects.

George Votsis graduated from the American College of Greece in 1992 and from the Department of Electrical and Computer Engineering, the National Technical University of Athens in 1997, where he is currently working toward the Ph.D. degree. His current research interests lie in the areas of image and video processing, machine vision, human—computer interaction and artificial neural networks. He is a member of the Technical Chamber of Greece. He has published three papers in international journals more than ten in proceedings of international conferences. Since 1997 he has participated in five research projects.

Stefanos Kollias obtained his Diploma from National Technical University of Athens (NTUA) in 1979, his M.Sc. in communication engineering in 1980 from UMIST, U.K., and his Ph.D. in signal processing from the Computer Science Division of NTUA in 1984. At that time he was given an IEEE ComSoc Scholarship. From 1987 to 1996 he has been Assistant and Associate Professor in the Electrical and Computer Engineering Department of NTUA. In 1987-1988 he was a visiting Research Scientist at Columbia University, New York. Since 1997 he has been Professor and Director of the Image, Video and Multimedia Systems Laboratory of NTUA. His research interests include image and video analysis, intelligent multimedia systems, artificial neural networks and hybrid systems. He has published more than 150 papers. He has been a member of the technical committee or invited speaker in 40 international conferences, reviewer of 25 journals and belongs to the editorial board of *Neural Networks*. He has been leading 40 R&D projects at European and National level, also being the coordinator of the TMR PHYSTA project.

Winfried Fellenz studied Computer Science and Electrical Engineering at Dortmund University from 1985, graduating in spring 1992. In the same year he was a Visiting Scientist at Brown University in Providence. He received his Dr.-Ing. from the University of Paderborn in 1997. After spending two years as an Associated Researcher at the Department of Computer Vision at TU Berlin, he joined the Department of Mathematics at King's College in London as a post-doc in spring 1998. His research interests include neural networks and learning, visual perception and active vision.

John G. Taylor was trained as a theoretical physicist in the Universities of London and Cambridge and had obtained various positions in Universities in the United Kingdom, United States, and Europe in physics and mathematics. He has created the center for Neural Networks at King's College, London, in 1990. He was appointed Emeritus Professor of Mathematics of London University in 1996 and Guest Scientist at the Research center in Juelich, Germany, 1996-1998, working on brain imaging and data analysis. He is presently European Editor-in-Chief of *Neural Networks* and was President of the International Neural Network Society (1995) and the European Neural Network Society (1993-1994). He has published over 450 scientific papers in theoretical physics, astronomy, particle physics, pure mathematics, neural networks, higher cognitive processes, brain imaging, and consciousness. He has authored 12 books and edited 13 others. His present research interests are: neural networks, industrial applications, dynamics of learning processes, stochastic neural chips and their applications and higher cognitive brain processes including consciousness.

References

[1] R. Cowie and E. Douglas-Cowie, "Speakers and hearers are people: Reflections on speech deterioration as a consequence of acquired deafness," in *Profound Deafness and Speech Communication,* K-E. Spens and G. Plant, Eds. London, UK: Whurr, 1995, pp. 510-527.

[2] H. Giles and P. Smith, "Accommodation theory: Optimal levels of convergence," in *Language and Social Psychology*, H. Giles and R. St. Clair, Eds. Oxford, UK: Blackwell, 1979, pp. 45-65.

[3] "Prosody and conversation," in *Language and Speech (Special Issue),* M. Sverts and J. Hirschman, Eds., vol. 41, no. 3-4, 1998.

[4] *DSM III—Diagnostic and Statistical Manual of Mental Disorders*, *III*. Washington, DC: American Psychiatric Association, 1987.

[5] S. McGilloway, R. Cowie, and E. Douglas-Cowie, "Prosodic signs of emotion in speech: Preliminary results from a new technique for automatic statistical analysis," in *Proc. XIIIth Int. Congr. Phonetic Sciences*, vol. 1. Stockholm, Sweden: 1995, pp. 250-253.

[6] J.A. Russell, "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies," *Psychol. Bull.*, vol. 115, pp. 102-141, 1994.

[7] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Proc. 4th Int. Conf. Spoken Language Processing*. Philadelphia, PA, 1996, pp. 1989-1992.

[8] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psych*, vol. 70, no. 3, pp. 614-636, 1996.

[9] E. Douglas-Cowie and R. Cowie, "International settings as markers of discourse units in telephone conversations," *Language and Speech (Special Issue*, *Prosody and Conversation)*, vol. 41, no. 3-4, pp. 351-374, 1998.

[10] I. Murray and J. Arnott, "Synthesizing emotions in speech: Is it time to get excited?," in *Proc. 4th Int. Conf. Spoken Language Processing*, Philadelphia, PA, 1996, pp. 1816-1819.

[11] M.A. Walker, J.E. Cahn, and S.J. Whittaker, "Improvising linguistic style: Social and affect bases for agent personality," in *Proc. Int. Conf. Autonomous Agents ACM SIGART*, New York, 1997, pp. 96-105.

[12] R. Cornelius, *The Science of Emotion*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[13] K. Oatley and P. Johnson-Laird, "Communicative theory of emotions: Empirical tests, mental models & implications for social interaction," in *Goals and Affect*, L. Martin and A. Tessler, Eds. Hillsdale, NJ: Erlbaum, 1995.

[14] R. Plutchik, *The Psychology and Biology of Emotion*. New York: Harper Collins, 1994, pp. 58.

[15] K.R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychological Bulletin,* vol. 99, pp. 143-165, 1986.

[16] K.R. Scherer, "On the nature and function of emotion: A component process approach," in *Approaches to Emotion,* K.R. Scherer and P. Ekman, Eds. Hillsdale, NJ: Erlbaum, 1984.

[17] K. Oatley and J.M. Jenkins, *Understanding Emotions.* Oxford, UK: Blackwell, 1996.

[18] M.B. Arnold, *Emotion and Personality: Vol 2. Physiological Aspects.* New York: Columbia Univ. Press, 1960.

[19] R.S. Lazarus, *Emotion and Adaptation*. New York: Oxford Univ. Press, 1991.

[20] K. Oatley and P. Johnson-Laird, "Towards a cognitive theory of emotions," *Cognition and Emotion*, vol. 1, pp. 29-50, 1987.

[21] P. Ekman, "Basic Emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. Power, Eds. Chichester, UK: Wiley, 1999.

[22] C.M. Whissel, "The dictionary of affect in language," *Emotion: Theory, Research and Experience: Vol. 4, The Measurement of Emotions*, R. Plutchik and H. Kellerman, Eds. New York: Academic, 1989.

[23] R. Plutchik, *Emotion: A Psychoevolutionary Synthesis*. New York: Harper & Row, 1980.

[24] J. Russell and J. Fernandez-Dols, *The Psychology of Facial Expression*. Cambridge, UK: Cambridge University Press, 1997.

[25] T. Dalgleish and M. Power, Eds., *Handbook of Cognition and Emotion*. Chichester, UK: Wiley, 1999.

[26] S.S. Tomkins, "Affect theory," in *Emotion in the Human Face*, P. Ekman, Ed., 2nd ed. New York: Cambridge Univ. Press, 1982.

[27] N.A. Fox, "If it's not left it's right," *Amer. Psychol.,* vol. 46, pp. 863-872, 1992.

[28] J.R. Averill, "Acquisition of emotions in adulthood," in *The Social Construction of Emotions*, R. Harre, Ed. Oxford, UK: Blackwell, 1986, pp. 100.

[29] H. Morsbach and W.J. Tyler, "A Japanese emotion: Amae," in *The Social Construction of Emotions*, R. Harre, Ed. Oxford, UK: Blackwell, 1986, pp. 289-307.

[30] R. Harre and R. Finlay-Jones, "Emotion talk across times," in *The Social Construction of Emotions*, R. Harre, Ed. Oxford, UK: Blackwell, 1986, pp. 220-233.

[31] M. Cowan, "Pitch and intensity characteristics of stage speech," *Arch. Speech*, suppl. to Dec. issue, 1936.

[32] G. Fairbanks and W. Pronovost, "An experimental study of the pitch characteristics of the voice during the expression of emotion," *Speech Monograph,* vol. 6, pp. 87-104, 1939.

[33] G.E. Lynch, "A phonophotographic study of trained and untrained voices reading factual and dramatic material," *Arch. Speech,* vol. 1, pp. 9-25, 1934.

[34] R. Frick, "Communicating emotion: The role of prosodic features," *Psychol. Bull.,* vol. 97, pp. 412-429, 1985.

[35] I.R. Murray and J.L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Amer*., vol. 93, no. 2, pp. 1097-1108, 1993.

[36] M. Schubiger, *English Intonation. Its Form and Function*. Tubingen, Germany: Niemeyer, 1958.

[37] D. Crystal, *Prosodic Systems and Intonation in English.* London, UK: Cambridge Univ. Press, 1969.

[38] D. Crystal, *The English Tone of Voice*. London, UK: Edward Arnold, 1975.

[39] J.D. O'Connor and G.F. Arnold, *Intonation of Colloquial English,* 2nd ed. London, UK: Longman, 1973.

[40] P. Lieberman and S.D. Michaels, "Some aspects of fundamental frequency, envelope amplitude and the emotional content of speech," *J. Acoust. Soc. Am*., vol. 34, pp. 922-927, 1962.

[41] E. Uldall, "Attitudinal meanings conveyed by intonational contours," *Language and Speech,* vol. 3, pp. 223-234, 1960.

[42] S. Ladd, K. Siverman, G. Bergmann, and K. Scherer, "Evidence for independent function of intonation contour type, voice quality, and F0 in signaling speaker affect," *J. Acoust. Soc. Am*., vol. 78, no. 2, pp. 435-444, 1985.

[43] K. Scherer, D.R. Ladd, and K. Silverman, "Vocal cues to speaker affect: Testing two models," *J. Acoust. Soc. Am*., vol. 76, pp. 1346-1356, 1984.

[44] E. Couper Kuhlen, *An Introduction to English Prosody*. London, UK: Edward Arnold, 1986, pp. 176.

[45] D.R. Ladd, *Intonational Phonology*. Cambridge, UK: Cambridge Univ. Press, 1996.

[46] J.R. Davitz and L.J. Davitz, "The communication of meanings by content-free speech," *J. Commun.,* vol. 9, pp. 6-13, 1959.

[47] G. Fairbanks and W. Pronovost, "Vocal pitch during simulated emotion," *Science,* vol. 88, pp. 382-383, 1938.

[48] G. Fairbanks, "Recent experimental investigations of pitch in speech," *J. Acoust. Soc. Am*., vol. 11, pp. 457-466, 1940.

[49] G. Fairbanks and L. Hoaglin, "An experimental study of the durational characteristics of the voice during the expression of emotion," *Speech Monograph,* vol. 8, pp. 85-91, 1941.

[50] J. Starkweather, "Content-free speech as a source of information about the speaker," *J. Abnorm. Soc. Psychol*., vol. 52, pp. 394-402, 1956.

[51] P. Roach, R. Stibbard, J. Osborne, S. Arnfield, and J. Setter, "Transcriptions of prosodic and paralinguistic features of emotional speech," *J. Int. Phonetic Assoc*., vol. 28, pp. 83-94.

[52] R. Cowie and E. Douglas-Cowie, *Postlingually Acquired Deafness: Speech Deterioration and the Wider Consequences*. Berlin, Germany: Mouton de Gruyter, 1992.

[53] S. McGilloway, R. Cowie, and E. Douglas-Cowie, "Prosodic signs of emotion in speech: Preliminary results from a new technique for automatic statistical analysis," in *Proc. 13th ICPhS*, Stockholm, Sweden, 1995, pp. 1989-1991.

[54] S. McGilloway, "Negative symptoms and speech parameters in schizophrenia," Ph.D. dissertation, Faculty of Medicine, Queen's University, Belfast, UK, 1997.

[55] G. Izzo, PHYSTA Project Report, "Multiresolution techniques and emotional speech," NTUA Image Processing Laboratory, Athens, 1998.

[56] EC TMR Project PHYSTA Report, "Hybrid systems for feature to symbol extraction" [Online]. Available: http://www.imge.ece.ntua.gr/physta/face-speech-features, July, 1998.

[57] E. Mousset, W.A. Ainsworth, and J. Fonollosa, "A comparison of recent several methods of fundamental frequency estimation," in *Proc. ICSLP 96*, Philadelphia, PA, pp. 1273-1276.

[58] J. Suzuki, M. Setoh, and T. Shimamura, "Extraction of precise fundamental frequency based on harmonic structure of speech," in *Proc. 15th Int. Congr. Acoustics*, vol. 3, 1995, pp. 161-164.

[59] M.P. Karnell, K.D. Hall, and K.L. Landahl, "Comparison of fundamental frequency and perturbation measurements among 3 analysis systems," *J. Voice,* vol. 9, pp. 383-393, 1995.

[60] B.L. McKinley and G.H. Whipple, "Model based speech pause detection," *IEEE Int. Conf. Acoust., Speech and Signal Processing,* 1997, pp. 1179-1182.

[61] H.F. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. ICSLP 96*, Philadelphia, PA, pp. 1261-1264.

[62] W. Reichl and G. Ruske, "Syllable segmentation of continuous speech with artificial neural networks," in *Proc. Eurospeech 93*, vol. 3. Berlin, Germany, pp. 1771-1774.

[63] Y. Bengio, R. De Mori, G. Flammia, and H. Kompe, "Phonetically motivated acoustic parameters for continuos speech recognition using neural networks," in *Proc. Eurospeech-91*, Genova, Italy, pp. 551-554.

[64] A. Esposito, C.E. Ezin, and M. Ceccarelli, "Preprocessing and neural classification of the English stops [b, d, g, p, t, k]," in *Proc. ICSLP 96*, vol. 2. Philadelphia, PA, pp. 1249-1252.

[65] A. Esposito and C. Ezin C., "A Rasta-PLP and TDNN based automatic system for recognizing stop consonants: Performance studies," Vietri sul Mare (SA), Italy, IIASS Int. Rep. I9602b, 1996.

[66] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge, UK: Cambridge Univ. Press, 1980.

[67] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of voice qualities," *Acta Otolaryngologica 90*, pp. 441-451, 1980.

[68] W. Ding and H. Kasuya, "A novel approach to the estimation of voice source and vocal tract parameters from speech signals," in *Proc. ICSLP 96*, Philadelphia, PA, pp. 1257-1260.

[69] G. Klasmeyer and W. Sendlmeier, "Objective voice parameters to characterise the emotional content in speech," in *Proc. 13th Int. Congr. Phonetic Sciences*, vol. 2. Stockholm, Sweden, 1995, pp. 182-185.

[70] M. Liberman and A. Prince, "On stress and linguistic rhythm," *Linguistic Inquiry,* vol. 8, pp. 249-336, 1977.

[71] E. Keller and B. Zellner, "A statistical timing model for French," in *Proc. 13th Int. Congr. Phonetic Sciences*, vol. 3. Stockholm, Sweden, 1995, pp. 302-305.

[72] G. McRoberts, M. Studdert-Kennedy, and D.P. Shankweiler, "Role of fundamental frequency in signalling linguistic stress and affect: Evidence for a dissociation," *Perception and Psychophysics* 57, pp. 159-174, 1995.

[73] A. Cutler and M. Pearson, "On the analysis of prosodic turn-taking cues," in *Intonation in Discourse,* C. Johns-Lewis, Ed. London, UK: Croom Helm, 1986, pp. 139-155.

[74] G. Brown, K. Currie, and J. Kenworthy, *Questions of Intonation*. London, UK: Croom Helm, 1980.

[75] V. Bruce and A. Young, *In the Eye of the Beholder: The Science of Face Perception*. London, UK: Oxford Univ. Press, 1998.

[76] C. Darwin, *The Expression of Emotions in Man and Animals*, John Murray, Ed., 1872. Reprinted by Univ. Chicago Press, 1965.

[77] P. Ekman, *Darwin and Facial Expressions.* New York: Academic, 1973.

[78] M. Davis and H. College, *Recognition of Facial Expressions*. New York: Arno Press, 1975.

[79] K. Scherer and P. Ekman, *Approaches to Emotion*. Mahwah, NJ: Lawrence Erlbaum Associates, 1984.

[80] A. Young and H. Ellis, *Handbook of Research on Face Processing*. Amsterdam, The Netherlands: Elsevier Science Publishers, 1989.

[81] P. Ekman, T. Huang, Y. Sejnowski, and J. Hager, "NSF planning workshop on facial expression understanding," National Science Foundation, Human Interaction Lab, Tech. Rep., 1992.

[82] V. Bruce, *Recognizing Faces*. Mahwah, NJ: Lawrence Erlbaum, 1988.

[83] J. Jaynes, *The Breakdown of the Bicameral Mind*. London, UK: Oxford Univ. Press, 1976.

[84] J. Gabrieli, R. Poldrack, and J. Desmond, "The role of the left prefontal cortex in language and memory," *Proc. Natl. Academy Sciences USA*, vol. 95, 1998, pp. 906-913.

[85] M.S. Bartlett, J.C. Hager, P. Ekman, and T.J. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, pp. 253-263, 1999.

[86] P. Ekman, J. Hager, and E. Rosenberg, "FACSAID: A computer database for predicting affective phenomena from facial movement," [Online]. Available: http:/www.nirc.com/facsaid.html

[87] R. Cabeza and J. Nyburg, *Cognitive Neuroscience,* vol. 9, 1997, pp. 1-26.

[88] J. Lien, T. Kanade, J. Cohn, and C. Li, "Subtly different facial expression recognition and emotion expression intensity estimation," in *Proc. IEEE CVPR,* Santa Barbara, CA, 1998, pp. 853-859.

[89] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. 7th Intl. Joint Conf. on AI*, 1981.

[90] M.L. Phillips, A.W. Young, and C. Senior, "A specific neural substrate for perceiving facial expressions of disgust," *Nature*, vol. 389, pp. 495-498, 1997.

[91] I. Essa and A. Pentland, "A vision system for observing and extracting facial action parameters," in *Proc. Int. Conf. CVPR*, 1994, pp. 76-83.

[92] S.K. Scott, A.W. Young, and A.J. Calder, "Impaired auditory recognition of fear and anger following bilateral amygdala lesions," *Nature,* vol. 385, pp. 254-257, 1997.

[93] O.S.P. Scalaidhe, F.A.W. Wilson, and P.S. Goldman Rakic, *Science*, vol. 278, pp. 1135-1108, 1997.

[94] N. Intrator, D. Reisfeld, and Y. Yeshrun, "Face recognition using a hybrid supervised/unsupervised neural network," *Patt. Recognit Lett.,* vol. 17, pp. 67-76, 1996.

[95] N. Arad and D. Reisfeld, "Image warping using few anchor points and radial functions," *Computer Graphics Forum*, vol. 14, no. 1, pp. 35-46, 1994.

[96] S. Shibui, H. Yamada, T. Sato, and K. Shigemasu, "Categorical perception and semantic information processing of facial expressions," *Perception*, vol. 28 S, pp. 114, 1999.

[97] *MPEG4 SNHC: Face and Body Definition and Animation Parameters*, ISO/IEC JTC1/SC29/WG11 MPEG96/N1365, 1996.

[98] P. Ekman and W. Friesen, *The Facial Action Coding System*. San Francisco, CA: Consulting Psychologists Press, 1978.

[99] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 6, pp. 569-579, 1993.

[100] P. Roivainen, H. Li, and R. Forcheimer, "3-D motion estimation in model-based facial image coding," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 545-555, 1993.

[101] K. Mase, "Recognition of facial expression from optical flow," *IEICE Trans.*, vol. E74, pp. 3474-3483, 1991.

[102] C. Pelachaud, N. Badler, and M. Viaud, "Final report to NSF of the standards for facial animation workshop," NSF, Univ. Penn., Philadelphia, PA, Tech. Rep., 1994.

[103] P. Ekman, T. Huang, T. Sejnowski, and J. Hager, Eds., "Final report to NSF of the planning workshop on facial expression understanding," NSF, Human Interaction Lab., UCSF, CA, Tech. Rep., 1993.

[104] P. Ekman and W. Friesen, *Unmasking the Face*. Englewood Cliffs, NJ: Prentice-Hall, 1975.

[105] C. Padgett and G.W. Cottrell, "Representing face images for emotion classification," *Advances in Neural Information Processing Systems,* vol. 9. Cambridge, MA: MIT Press, 1997, pp. 894.

[106] C. Padgett, G. Cottrell, and B. Adolps, "Categorical perception in facial emotion classification," in *Proc. Cognitive Science Conf.,* vol. 18, 1996, pp. 249-253.

[107] B.A. Golomb, D.T. Lawrence, and T.J. Sejnowski, "Sexnet: A neural net identifies sex from human faces," in *NIPS 3*, R.P. Lippman, J. Moody, and D.S. Touretzky, Eds. San Francisco, CA: Morgan Kaufmann, 1991, pp. 572-577.

[108] J.N. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *J. Personality Social Psychol.*, vol. 37, pp. 2049-2059, 1979.

[109] B. Horn and B. Schunk, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185-203, 1981.

[110] D. Heeger, "Optical flow using spatiotemporal filters," *Int. J. Comput. Vision,* vol. 1, pp. 279-302, 1988.

[111] M. Abdel-Mottaleb, R. Chellappa, and A. Rosenfeld, "Binocular motion stereo using MAP estimation," in *IEEE Conf. Computer Vision and Pattern Recognition*, 1993, pp. 321-327.

[112] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int. J. Comput. Vision,* vol. 2, pp. 283-310, 1989.

[113] Y. Yacoob and L. Davis, "Computing spatio-temporal representations of human faces," in *Proc. Computer Vision and Pattern Recognition Conf.,* 1994, pp. 70-75.

[114] N. Tsapatsoulis, I. Avrithis, and S. Kollias, "On the use of radon transform for facial expression recognition," in *Proc. 5th Intl. Conf. Information Systems Analysis and Synthesis*, Orlando, FL, Jul. 1999.

[115] Y. Yacoob and L.S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Trans. Pattern Anal. Machine Intell.* vol. 18, no. 6, pp. 636-642, 1996.

[116] M.J. Black and P. Anandan, "The robust estimation of optical flow," in *Proc. Int. Conf. Computer Vision*, Berlin, Germany, 1993, pp. 231-236.

[117] Y.T. Wu, T. Kanade, J. Cohn, and C.C. Li, "Optical flow estimation using wavlet motion model," in *Proc. Int. Conf. Computer Vision (ICCV'98)*, Bombay, India, Jan. 1998.

[118] T. Sakaguchi, J. Ohya, and F. Kishino, "Facial expression recognition from image sequence using hidden Markov model," *VLBV* 95, A-5, 1995.

[119] T. Otsuka and J. Ohya, "Recognition of facial expressions using HMM with continuous output probabilities," in *Proc. 5th IEEE Int. Workshop on Robot and Human Communication RO-MAN*, 1996, pp. 323-328.

[120] T. Otsuka and J. Ohya, "Recognizing multiple persons' facial expressions using HMM based on automatic extraction of frames from image sequences," in *Proc. IEEE Int. Conf. on Image Proc.*, vol. 2, 1997, pp. 546-549.

[121] Y. Yacoob and L.S. Davis, "Recognizing human facial expressions," in *Proc. 2nd Workshop on Visual Form*, Capri, Italy, 1994, pp. 584-593.

[122] M. Rosenblum, Y. Yacoob, and L. Davis, "Human emotion recognition from motion using a radial basis function network architecture," *IEEE Trans. Neural Networks,* vol. 7, no. 5, pp. 1121 -1138, 1996.

[123] N. Thalmann, P. Kalra, and M. Escher, "Face to virtual face," *Proc. IEEE*, vol. 86, pp. 870-883, 1998.

[124] I. Essa and A. Pentland, "Coding, analysis, interpretation and recognition of facial expressions," MIT Media Lab., Cambridge, MA, Tech. Rep. 325, 1995.

[125] I. Essa, T. Darrell, and A. Pentland, "Tracking facial motion," in *Proc. Workshop on Motion of Nonrigid and Articulated Objects*, 1994, pp. 36-42.

[126] I. Essa, S. Sclaroff, and A. Pentland, "Physically-based modeling for graphics and vision," in *Directions in Geometric Computing. Information Geometers*, R. Martin, Ed. U.K., 1993.

[127] S.M. Platt and N.L. Badler, "Animating facial expression," in *Proc. ACM SIGGRAPH Conference,* vol. 15, no. 3, 1981, pp. 245-252.

[128] N. Tsapatsoulis, M. Leonidou, and S. Kollias, "Facial expression recognition using HMM with observation dependent transition matrix," in *Proc. MMSP'98,* Portofino CA, Dec. 1998.

[129] S. McKenna and S. Gong, "Tracking faces," in *Proc. 2nd Intl. Conf. on Automatic Face and Gesture Recognition*, 1996, pp. 271-276.

[130] J. Crowley and F. Berard, "Multi-modal tracking of faces for video communications," in *Proc. IEEE CVPR*, Puerto Rico, June 1997, pp. 640-645.

[131] H. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, "Multi-modal system for locating heads and faces," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Vermont, Oct. 1996, pp. 88-93.

[132] M. Collobert, et al., "Listen: A system for locating and tracking individual speakers," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Vermont, Oct. 1996, pp. 283-288.

[133] M. Hunke and A. Waibel, "Face locating and tracking for human computer interaction," *IEEE Computer*, pp. 1277-1281, Nov. 1994.

[134] S. Basu, I. Essa, and A. Pentland, "Motion regulization for model-based head tracking," in *Proc. 13th Int. Conf. on Pattern Recognition*, Aug. 1996.

[135] P. Fieguth and D. Terzopoulos, "Color-based tracking of image regions with changes in geometry and illumination," in *Proc. IEEE CVPR*, 1996, pp. 403-410.

[136] J. Terillon, M. David, and S. Akamatsu, "Automatic face detection in natural scene images using a skin color model and moments," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Nara, Japan, Apr. 1998.

[137] Y. Raja, S. McKenna, and S. Gong, "Tracking and segmenting people in varying lighting conditions using color," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, Nara, Japan, Apr. 1998, pp. 228-233.

[138] T. Maurer and V.D.C. Malsburg, "Tracking and learning graphs on image sequences of faces," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Vermont, Oct. 1996, pp. 176-181.

[139] V. Kruger and G. Sommer, "Affine face tracking using a wavelet network," in *Proc. Int. Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, Corfu, Greece, Sept. 1999.

[140] S. Gong, S. McKenna, and S. Collins, "An investigation into face pose distributions," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Vermont, 1996, pp. 265-270.

[141] A. Zelinsky and J. Heinzmaan, "Real-time visual recognition of facial gestures for HCI," in *Proc. Int. Conf. Automatic Face and Gesture Recognition,* Oct. 1996, pp. 351-356.

[142] J.F. Cohn, A. Zlochower, J. Lien, and T. Kanade, "Automated face analysis by feature point tracking has high concurrent validity with manual FACS coding," *Psychophysiology*, vol. 26, pp. 35-43, 1999.

[143] H. Wu, T. Yokoyama, D. Pramadihanto, and M. Yachida, "Face and facial feature extraction from color images," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Oct. 1996, pp. 345-350.

[144] R. Herpers, M. Michaelis, K.H. Lichtenauer, and G. Sommer, "Edge and keypoint detection in facial regions," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Vermont, Oct. 1996, pp. 212-217.

[145] R. Herpers, H. Kattner, H. Rodax, and G. Sommer, "An attentional processing strategy to detect and analyse the prominent facial regions," in *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, Zurich, Switzerland, 1995, pp. 214-220.

[146] M. Black, Y. Yacoob, A. Jepson, and D. Fleet, "Learning parameterized models of image motion," in *Proc. IEEE CVPR*, 1997, pp. 561-567.

[147] H. Wang and S. Chang, "A highly efficient system for automatic face region detection in MPEG video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, pp. 615-628, 1997.

[148] F. Parke and K. Waters, *Computer Facial Animation*, A.K. Peters, Ed. Wellesley, MA: A.K. Peters, 1996.

[149] K. Lam and H. Yan, "An analytic to holistic approach for face recognition based on a single frontal view," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 20, no. 7, July 1998.

[150] EC TMR Project PHYSTA Report, "Confidence estimation and on-line retraining of neural networks" [Online]. Available: http://www.image.ece.ntua.gr/physta/confidence, Sept. 1999.

[151] EC TMR Project PHYSTA, "Test material format and availability" [Online]. Available: http://www.image.ece.ntua.gr/physta/testmaterial, Jan. 1999.

[152] P. Ekman and W. Friesen, *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists' Press, 1975.

[153] A. Savage and R. Cowie, "Are artificial neural nets as smart as a rat?," *Network,* vol. 3, pp. 47-59, 1992.

[154] M.L. Ginsberg, *Readings in Nonmonotonic Reasoning*. San Mateo, CA: Morgan Kaufmann, 1987.

[155] C.E. Osgood, G.J. Suci, and P.H. Tannenbaum, *The Measurement of Meaning*. Urbana, IL: Univ. of Illinois Press, 1957.

[156] J.R. Davitz, Ed., *The Communication of Emotional Meaning*. New York: McGraw-Hill, 1964.

[157] I. Fonagy, "A new method of investigating the perception of prosodic features," *Language and Speech*, vol. 21, pp. 34-49, 1978.

[158] C.E. Williams and K.N. Stevens, "On determining the emotional state of pilots during flight: An exploratory study," *Aerospace Medicine*, vol. 40, pp. 1369-1372, 1969.

[159] W.L. Hoffe, "Ueber beziehung von sprachmelodie und lautstarke [On the relation between speech melody and intensity]," *Phonetica*, vol. 5, pp. 129-159, 1960.

[160] Z. Havrdova and M. Moravek, "Changes of the voice expression during suggestively influenced states of experiencing," *Activitas Nervosa Superior*, vol. 21, pp. 33-35, 1979.

[161] R. Van Bezooijen, *Characteristics and Recognizability of Vocal Expressions of Emotions*. Dordrecht, The Netherlands: Foris, 1984.

[162] G. Kotlyar and V. Mozorov, "Acoustic correlates of the emotional content of vocalized speech, *J. Acoust. Academy of Sciences of the USSR*, vol. 22, pp. 208-211, 1976.

[163] A. Muller, "Experimentelle untersuchungen zur stimmlichen darstellung von gefuehlen [Experimental studies on vocal portrayal of emotion]," Ph.D. dissertation, Univ. Gottingen, Germany, 1960.

[164] I. Fonagy, "Emotions, voice and music," in *Language and Speech*, J. Sundberg, Ed., vol. 21, pp. 34-49, 1978.

[165] C.E. Williams and K.N. Stevens, "Emotions and speech: Some acoustic correlates," *J. Acoust. Soc. Am.*, vol. 52, pp. 1238-1250, 1972.

[166] I. Fonagy and K. Magdics, "Emotional patterns in intonation and music," *Z. Phonet. Sprachwiss. Kommunikationsforsch*, vol. 16, pp. 293-326, 1963.

[167] F. Trojan, *Der Ausdruck der Sprechstimme.* Wien-Dusseldorf, Germany: W. Maudrich, 1952.

[168] A. Oster and A. Risberg, "The identification of the mood of a speaker by hearing impaired listeners," *Speech Transmission Lab. Quarterly Progress Status Report 4*, Stockholm, pp. 79-90, 1986.

[169] K. Sedlacek and A. Sychra, "Die melodie als faktor des emotionellen ausdrucks [Speech melody as a means of emotional expression]," *Folia Phoniatrica,* vol. 15, pp. 89-98, 1963.

[170] G.L. Huttar, "Relations between prosodic variables and emotions in normal American English utterances," *J. Speech and Hearing Research,* vol. 11, pp. 481-487, 1968.

[171] R. Coleman and R. Williams, "Identification of emotional states using perceptual and acoustic analyses," in *Care of the Professional Voice*, vol. 1 , V. Lawrence and B. Weinberg, Eds. New York: The Voice Foundation, 1979.

[172] L. Kaiser, "Communication of affects by single vowels," *Synthese,* vol. 14, pp. 300-319, 1962.

[173] W. Johnson, R. Emde, K. Scherer, and M. Klinnert, "Recognition of emotion from vocal cues," *Arch. Gen. Psych.*, vol. 43, pp. 280-283, 1986.

[174] W. Hargreaves, J. Starkweather, and K. Blacker, "Voice quality in depression," *J. Ab. Psych.,* vol. 7, pp. 218-220, 1965.

[175] N. Utsuki and N. Okamura, "Relationship between emotional state and fundamental frequency of speech," *Rep. Aeromedical Laboratory, Japan Air Self-Defense Force,* vol. 16, pp. 179-188, 1976.

[176] J. Sulc, "Emotional changes in human voice," *Activitas Nervosa Superior*, vol. 19, pp. 215-216, 1977.

[177] F. Costanzo, N. Markel, and P. Costanzo, "Voice quality profile and perceived emotion," *J. Counseling Psych.,* vol. 16, pp. 267-270, 1969.

[178] M.A.K. Halliday, *Intonation and Grammar in British English*. The Hague, The Netherlands: Mouton, 1967.

[179] P. Tench *The Intonation Systems of English*. London, UK: Cassell, 1996.

[180] J.W. Hicks, "An acoustical/temporal analysis of emotional stress in speech," Ph.D. dissertation, Univ. Florida, *Dissertation Abstracts International,* vol. 41, 4A, 1978.

[181] J. Jaffe and S. Feldstein, *Rhythms of Dialogue*. New York: Academic Press, 1960.

[182] E.T. Rolls, *The Brain and Emotion*. Oxford, UK: Oxford University Press, 1999.

[183] R. Cowie, E. Douglas-Cowie, and A. Romano, "Changing emotional tone in dialogue and its prosodic correlates," in *Proc. ESCA Workshop on Dialogue and Prosody*, Eindhoven, The Netherlands, 1999.

[184] M. Fishbein and I. Ajazen, *Belief, Attitude, Intention and Behavior: An Introduction to Theory and Research*. Reading, MA: Addison-Wesley, 1975.

[185] S.J. Breckler and E.C. Wiggins, "On defining attitude theory: Once more with feeling," in *Attitude Structure and Function,* A.R. Pratakanis, S.J Breckler, and A.G. Greenwald, Eds. Hillsdale, NJ: Erlbaum, 1989.

[186] C. Moore, J. Cohn, and G. Katz, "Quantitative description and differentiation of fundamental frequency contours," *Comput. Speech and Language*, vol. 8, pp. 385-404, 1994.

[187] K. Scherer, "nonlinguistic indicators of emotion and psychopathology," in *Emotions in Personality and Psychopathology*, C.E. Izard, Ed. New York: Plenum, 1979.

[188] A. Zlochower and J. Cohn, "Vocal timing in face-to-face interactions of clinically depressed and nondepressed mothers and the 4-month-old infants," *Infant Behavior and Development*, vol. 19, pp. 373-376, 1996.

[189] A. Nilsonne, "Speech characteristics as indicators of depressive illness," *Acta Psychiatrica Scandinavica*, vol. 77, pp. 253-263, 1988.

[190] R. Cowie, A. Wichmann, E. Douglas-Cowie, P. Hartley, and C. Smith, "The prosodic correlates of expressive reading," in *Proc. 14th Int. Congress of Phonetic Sciences*, San Francisco, CA, 1999, pp. 2327-2330.

[191] M. Kamachi, S. Yoshikawa, J. Gyoba, and S. Akamatsu, "The dynamics of facial expression judgment," *Perception*, vol. 28 S, pp. 54-55, 1999.

[192] W.E. Rinn, "The neuropsychology of facial expression: A review of neurological and psychological mechanisms for producing facial expressions," *Psychological Bulletin*, vol. 95, pp. 52-77, 1984.

[193] I.J. Roseman, "Cognitive determinants of emotion," in *Review of Personality and Social Psychology: Vol. 5 Emotions, Relationships, and Health*, P. Shaver, Ed. Beverley Hills, CA: Sage, 1984.

[194] I.J. Roseman, M.S. Spindel, and P.E. Jose, "Appraisals of emotion-eliciting events: Testing a theory of discrete emotions," *J. Personality Social Psychol.*, vol. 59, pp. 899-915, 1990.

[195] A. Ortony, G.L. Clore, and A. Collins, *The Cognitive Structure of Emotions.* Cambridge, UK: Cambridge Univ. Press, 1988.

[196] N.H. Frijda, *The Emotions*. Cambridge, UK: Cambridge Univ. Press, 1986.

[197] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, and W. Fellenz, "What a neural net needs to know about emotion words," in *Proc. 3rd World Multiconf. on Circuits, Systems, Comms. and Computers*, Athens, Greece, July 1999.

[198] L. Saxe, D. Dougherty, and T. Cross, "The validity of polygraph testing: Scientific analysis and public controversy," *Amer Psychol.,* vol. 40, pp. 355-366, 1985.

[199] A.F. Ax, "The psychological differentiation between fear and anger in humans," *Psychol. Med.,* vol. 15, pp. 433-442, 1953.

[200] K. Silverman, et al., "ToBI: A standard for labelling English prosody," in *Proc. Int. Conf. on Spoken Language Processing*, Banff, Canada, pp. 286-290, 1992.

[201] G. Katz, J. Cohn, and C. Moore, "A combination of vocal F0 dynamic and summary features discriminates between pragmatic categories of infant-directed speech," *Child Development,* vol. 67, pp. 205-217, 1996.

[202] J.E. Cahn, "The generation of affect in synthesised speech," *J. American Voice I/O Society*, vol. 8, pp. 1-19, 1990.

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

Zhihong Zeng, *Member, IEEE*, 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 behaviour differs in visual appearance, audio profile, and timing from spontaneously occurring behaviour. To address this problem, efforts to develop algorithms that can process naturally occurring human affective behaviour 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 multi-cue 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 to 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, introductory and 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 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 that 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. 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 or 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 system of Lisetti and Nasoz [85], which combines facial expression and physiological signals to recognize the user's emotion like fear and anger and then to adapt an animated interface agent to mirror the user's emotion, 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, the proactive HCI tool of Maat and Pantic [89] capable of learning and analyzing the user's context-dependent behavioral patterns from multi-sensory data and of adapting the interaction accordingly, the automated Learning Companion of Kapoor et al. [72] that 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 need help, and the multimodal computer-aided learning system[1] at Beckman Institute UIUC where the computer avatar offers an appropriate tutoring strategy based on the information of user's facial expression, keywords, eye movement and task state. These systems represent initial efforts towards 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 system, and game and entertainment industry. These systems will change the ways

- *Zhihong Zeng and Thomas S. Huang are with Beckman Institute, University of Illinois at Urbana-Champaign. 405 N Mathews Av., Urbana, 61801. E-mail: {zhzeng,huang}@ifp.uiuc.edu..*
- *Maja Pantic is with Imperial College London, Department of Computing, 180 Queen's Gate, London SW7 2AZ, UK, and with University of Twente, Faculty of EEMCS, the Netherlands. E-mail: m.pantic@imperial.ac.uk.*
- *Glenn I. Roisman is with Psychology Department, University of Illinois at Urbana-Champaign. 603 East Daniel St., Champaign, IL 61820. E-mail: roisman@uiuc.edu.*

Manuscript received (August 12, 2007) and revised (December 29,2007).

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

in which we interact with computer systems. For example, an automatic service call center with an affect detector would be able to make 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 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 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 past 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 from 1972 [145]. Since late 90s, 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] from 1991. Early efforts toward 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 machine analysis of affective expressions, readers are referred to [115], [31], [102], [49], [96], [105], [130], [121], [98] that were published in 1992 to 2007 respectively.

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

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

Accordingly, reviewing the efforts toward 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 that time when these surveys were written, most of the available datasets 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, 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 behaviour differs in visual appearance, audio profile, and timing from spontaneously occurring behaviour. Due to this criticism received from both cognitive and computer scientists, the focus of the research in the field started to shift to automatic analysis of spontaneously displayed affective behavior. Several studies have recently emerged on machine analysis of spontaneous facial expressions (e.g., [10], [28], [135], [4]) and vocal expressions (e.g., [12], [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 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 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], [151]).

This paper introduces and surveys these recent advances in the research on human affect recognition. In contrast to previously published survey papers in the field, it focuses on the approaches which 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. The paper is organized as follows. Section 2 describes 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 the paper.

2 HUMAN AFFECT (EMOTION) PERCEPTION

Automatic affect recognition is inherently a multi-disciplinary 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 longstanding 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], [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 supported especially by the cross-cultural studies conducted by Ekman [40], [42] indicating that humans perceive certain basic emotions with respect to facial expression in the same way regardless of culture. This influence of basic emotion theory has resulted in the fact that most of 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 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 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., Feeltrace system [30]). In automatic emotion recognition systems that are based on the 2D dimensional emotion representation (e.g., [17] , [53]), the problem is often further simplified to 2-class (positive vs. negative and active vs. passive) or 4-class (quadrants of 2D space) classification.

One of the most influential emotion theories in modern psychology is the appraisal-based approach [117] that can be regarded as the extension of the dimensional approach described above. In this representation, 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 the 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 linguistists 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 while some studies (e.g., [1], [92]) indicated that a facial expression in the visual channel is the most important affective cue and correlates well with body as well as voice. Many studies have theoretically and empirically demonstrated the advantage of 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 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 objective description, leaving the inference about the conveyed message to high-level decision making. As indicated by Cohn [26], most commonly used sign judgment method used for 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 recognition of basic emotions according to Emotional FACS (EMFACS) rules[2], recognition of various affective states according to

FACS Affect Interpretation Database (FACSAID)[2] introduced by Ekman et al. [43], as well as for recognition of other complex psychological states such as depression [47] or pain [144]. AUs of the FACS are very suitable to be used in studies on human naturalistic facial behavior as the thousands of anatomically possible facial expressions (independently 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], [134]).

Speech is another important communicative modality in human-human interaction. Speech conveys affective information through explicit (linguistic) messages, and implicit (paralinguistic) messages that reflect the way 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 to analyze 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 non-basic 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 a 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], [113]). The human judgment agreement is typically higher for facial expression modality than it is 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 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], [117]). For example, it has been shown that temporal dynamics of facial behavior represents a critical factor for distinction between spontaneous and posed facial behavior (e.g., [28], [47], [135], [134]) as well as for categorization of complex behaviors like pain, shame, and amusement (e.g., [47], [144], [4], [87]). Based on these findings, we may expect that temporal dynamics of each modality separately (facial and vocal) and temporal correlations between the two modalities play an important role in 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 – where the expresser is (e.g., inside, on the street, in the car), what his or her 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 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 limitation of our knowledge and page, 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: Samal and Iyengar 1992 [115], Pantic and Rothkrantz 2000 [101], and Fasel and Luttin 2003, [49],
- surveys of techniques for automatic facial muscle action recognition and facial expression analysis: Tian et al. 2005 [130], and Pantic and Bartlett 2007 [98], and
- overviews of multimodal affect recognition methods: Cowie et al. 2001 [31], Pantic and Rothkrantz 2003 [102], Pantic et al. 2005 [105], Sebe et al. 2005 [121], Jaimes and Sebe 2005 [68], and Zeng et al. 2007 [152] (this is a short, preliminary version of the survey presented in the 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 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 and short lived,

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

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 ground truth is very time consuming, error prone, and expensive. This state of affairs makes automatic analysis of spontaneous emotional expression a very difficult task. Due to these difficulties, most of the existing studies on 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 microphone.

However, increasing evidence suggests that deliberate behaviour differs in visual appearance, audio profile, and timing from spontaneously occurring behaviour. 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 and 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 time than posed smiles (e.g., [28], [47], [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 behaviours usually fail to generalize to the subtlety and complexity of spontaneous affective behaviour.

In addition, most of the current human affect recognizers are evaluated using clear and constrained input (e.g., high quality visual and audio recording, non-occluded, and front-view 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 human-computer interaction are non-basic 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], [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 development of datasets that could be used for training and test of automatic systems for human affect analysis. Table 1 lists some noteworthy audio, visual, and audio-visual data resources that were reported in the literature. For each database, we provide the following information: affect elicitation method (i.e., whether the elicited affective displays are posed or spontaneous), size (the number of subjects and available data samples), modality (audio and/or visual), affect description (category or dimension), labeling scheme, and public accessibility. For other surveys of existing databases of human affective behavior, the readers are referred to [32], [59], [106].

As far as the databases of deliberate affective behavior are concerned, the following databases needs 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 non-basic affective states (six prototypical emotions, uncertainty, anxiety, boredom, and neutral). The MMI facial expression database [106], [98] is to our knowledge the most comprehensive dataset 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 view 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 datasets 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 databases containing facial and vocal deliberate displays of basic emotions and 4 cognitive states (interest, puzzlement, frustration and boredom).

The existing datasets of spontaneous affective behavior were collected in one of the following scenarios: human-human conversation, human-computer interaction, and use of a video kiosk. Human-human conversation scenarios include face-to-face interviews (e.g., [10], [38], [111], [65]), phone conversations (e.g., [34]), and meetings (e.g., [15], AMI[6]). Human computer interaction scenarios include Wizard of OZ scenarios (e.g., [13], SAL[7]), and computer-based dialogue systems (e.g., [83], [86]). In the video kiosk settings (e.g., [95], [98], [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 vs. negative (e.g., [15], [83]), dimensional descriptions in the evaluation-activation space (e.g., SAL[7]), and some application-dependent affective states are usually

[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

used as the data labels. Interest, boredom, confusion, frustration, fatigue, empathy, stress, irony, annoyance, amusement, helplessness, panic, shame, reprehension, and rebelliousness are some typical examples of the used application-dependent affect-interpretative labels (e.g., [95], [63], [13], [111])..

TABLE 1. AUDIO AND/OR VISUAL DATABASES OF HUMAN AFFECTIVE BEHAVIOR

Legend: A – audio, V – video, AV- audiovisual, N/A – not available, Y – yes, N – not yet

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 [119]	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	FEELTRACE	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	FEELTRACE	Y

TABLE 2. VISION-BASED AFFECT RECOGNITION

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

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	?	Vi: 60(MMI) 59(CK),70(OD)	Vi: 90.7
Valstar et al. 07 [134]	same as in [103]	GentleSVM-sigmoid	S, P	?	H, B	N	2	MMI: ?	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)

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 few dozens of AUs. Hence, the labeling schemes used to code data include FACS AUs (e.g., [10], [71], [106], [98], [111]), Feeltrace system for evaluation-activation dimensional description (e.g., [38], SAL[7]), self-report (e.g., [123], [65]), and human-observer judgment (e.g., [13], [15], [83], [95], [98]).

The current situation of emotion database research is considerably different from what was described in the comprehensive surveys written by Pantic and Rothkrantz in 2003 [102] and Cowie et al. in 2001 [31]. The current state of the art is advanced and can be summarized as follows (Table 1):

- 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 profile-view is shared on Internet [106], [98],
- 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 datasets 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 analysis of exaggerated expressions of basic emotions to 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 vision-based affect recognition studies focus on facial expression analysis. We can distinguish two main streams in the current research on machine analysis of facial expressions [26], [98]: recognition of affect and recognition of facial muscle action (facial action units). As explained above, facial action units are 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 non-basic affective states from deliberately displayed facial expressions including fatigue [60], [69], and mental states like agreeing, concentrated, disagreeing, interested, thinking, confused and frustration (e.g., [48], [72], [73], [129], [147]).

Most of the existing works on automatic facial expression recognition are based on deliberate and often exaggerated facial displays (e.g., [130]). However, several efforts have been recently reported on automatic analysis of spontaneous facial expression data (e.g., [9], [10], [11], [27], [28], [67], [88], [123], [135], [149], [87], [4], [134]). Some of them study automatic recognition of AUs rather than emotions from spontaneous facial displays (e.g., [9], [10], [11], [27], [28], [135], [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 machine analysis of spontaneous facial behavior were conducted using the datasets listed in Table 1 (e.g., [10], [149], [134]). For other studies new datasets were collected. Overall, the utilized data were collected in the following data-elicitation scenarios: human-human conversation (e.g., [10], [11], [28], [135], [149], [4]), Wizard of OZ scenario (e.g., [67]), or TV broadcast (e.g., [147]). Studies reported in [123], [147] explored automatic recognition of a subset of basic emotional expressions. The study of Zeng et al. [149] investigated separating emotional state from non-emotional states during the Adult Attachment Interview. Studies on separating posed from genuine smiles were reported in [28] and [134] and studies on 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 spatio-temporal 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 and her colleagues [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 Candide grid. Typical examples of appearance-feature-based methods are those of Bartlett et al. [9], [10], [11], [87], and Guo and Dyer [64], who used Gabor wavelets, of Whitehill and Omlin [143] who used Haar features, of Anderson and McOwen [2], who used a holistic spatial ratio face template, of Valstar et al. [136], who used temporal templates, and of 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 to design automatic facial expression recognizer. Typical examples of hybrid, geometric- and appearance-feature-based methods, are those proposed by Tian et al. (e.g., [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 uses 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 analysis of facial expressions under a small range of head motions. Thus, most of these methods focus on 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 of the face.

TABLE 3. AUDIO-BASED AFFECT RECOGNITION

Legend: 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 – number of classes, sub – number of subjects, samp – sample size (number of utterances), acc – accuracy, ? – missing entry, BL – Baseline, EER – equal error rate, NPN – negative/neutral/positive, NnN – Negative/non-negative, EnE – emotional/non-emotional , M – male, F – female, A – actor data, R – reading data, W – data of Wizard of OZ. ACC, AIBO, CSC and ISL are the database names listed in Table 1, OD – other database.

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	

TABLE 4. AUDIOVISUAL AFFECT RECOGNITION

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

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, prosody	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 languages
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	

A few approaches to automatic facial expression analysis are based on 3D face models. Huang and his colleagues (i.e., [25], [123], [141], [149]) used features extracted by a 3D face tracker called Piecewise Bezier Volume Deformation Tracker [128]. Cohn et al. [27] focused on analysis of brow action units and head movement based on a cylindrical head model [146]. Chang et al. [19] 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 multi-linear 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], [134]), the fusion of facial expression and body gesture (e.g., [7], [61], [62], [134]), and the fusion of facial expressions and postures from a sensor chair (e.g., [72], [73]), with the aim at improvement of affect recognition performance.

Finally, virtually all present approaches to automatic facial expression analysis are context insensitive. Exceptions from 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 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, and

environment, 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 type of the utilized data (spontaneous or posed, number of different subjects; sample size), whether the system is person-dependent/ independent, whether it performs in real time condition, what is the number of target classification categories, whether and which other cues besides the face have been used in the classification (head/ body/ eye/ posture/ task state/ other context), whether the system processes still images or videos, and 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 in machine analysis of facial affect has seen a lot of progress when compared to that described in the survey paper of Pantic and Rothkrantz from 2003 [102]. The current state of the art in the field is as follows:

- Methods have been proposed to detect attitudinal and non-basic affective states such as confusion, boredom, agreement, fatigue, frustration, and pain from facial expressions (e.g., [69], [72], [129], [147], [87]).
- Initial efforts were conducted to analyze and automatically discern posed (deliberate) facial displays from genuine (spontaneous) displays (e.g., [135], [134]).
- First attempts are reported towards vision-based analysis of spontaneous human behavior based on 3D face models (e.g., [123], [149]), based on fusing the information from facial expressions and head gestures (e.g., [27], [134]), and based on fusing the information from facial expressions and body gestures (e.g., [61]).
- Few attempts have been also made towards context-dependent interpretation of the observed facial behavior (e.g., [50], [69], [72], [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 in vocal affect recognition is also largely influenced by basic emotion theory. In turn, most of the existing efforts in this direction aim at recognition of a subset of basic emotions from speech signals. However, a few tentative studies were published recently on interpretation of speech signals in terms of certain application-dependent affective states. These studies are those of Hirschberg et al. [65] and Graciarena et al. [57], who attempted deception detection, of Liscombe et al. [84], who focused on detecting certainness, of Kwon et al. [79], who reported on stress detection, of Zhang et al. [157], who investigated speech-based analysis of confidence, confusion, and frustration, of Batliner et al. [12], who aimed at detecting trouble, of Ang et al. [3], who explored speech-based recognition of annoyance and frustration, and of Steidl et al. [125], who conducted studies on detection of empathy. In addition, few efforts towards 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 in machine analysis of human affects since recent studies in cognitive sciences showed that listeners seem to be rather accurate in decoding some non-basic affective states such as distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, and yawns (e.g., [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], [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], [94]), meetings (e.g., [94]), wizard of OZ scenarios (e.g., [12]), interview (e.g., [65]), and other dialogue systems (e.g., [14], [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], [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 was summarized in [31]. The popular features are prosodic features (e.g., pitch-related feature, energy-related features, speech rate), and spectral features (e.g., MFCC, cepstral features). Many studies show that pitch and energy among these features contribute 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 towards analysis of spontaneous human behavior, analysis of acoustic information

only will not 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 Litman et al. [86] and Schuller et al. [120], who used spoken words and acoustic features, of Lee and Narayanan [83], who used prosodic features, spoken words and information of repetition, of Graciarena et al. [57], who combined prosodic, lexical and cepstral features, and of 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] investigated also the role of the context information (e.g. subject, gender, 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 of language, discourse and context, automatic extraction of these related features is a difficult problem. First, existing automatic speech recognition 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 specifiy 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 non-basic affective states, including coarse affective states such as negative and non-negative states (e.g., [83]), application-dependent affective states (e.g., [3], [12], [65], [79], [157], [125]), and nonlinguistic vocalizations like laughter, cry (e.g., [133], [91], [97]).
- A few efforts have been made to integrate paralinguistic features and linguistic features such as lexical, dialogic, and discourse feature (e.g., [12], [35], [57], [83], [86], [120]).
- Few investigations have been conducted to make use of contextual information to improve the affect recognition performance (e.g., [52], [86]).
- Few studies have been reported to recognize the affective states across languages (e.g., [94], [133]).
- Some studies have investigated influence of ambiguity of human labeling on recognition performance (e.g., [3] [86]), and proposed measures to compare 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], [35]).

3.4 Audiovisual Affect Recognition

In the survey of Pantic and Rothkrantz in 2003, [102], only four studies were found that were focused 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 audio-visual affect recognition studies investigated recognition of the basic emotions from deliberate displays. Relatively few efforts have been reported toward detection of non-basic 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 4 cognitive states (interest, puzzlement, frustration and boredom) considering the importance of these cognitive states in human computer interaction. Related studies conducted on naturalistic data include that of Pal et al. [97], who designed a system to detect 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], [155]). Recently a few exceptional studies have been reported toward audiovisual affect analysis in spontaneous affect displays (e.g., [17], [53], [74], [97], [151], [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], Petridis and Pantic [108] used the recordings of people engaged in meetings (AMI corpus[6]), while 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 recognition of fine-grained affective states (e.g., basic emotions), recognition of coarse affective states was attempted in most of the aforementioned studies. Studies of Zeng et al. focus on audiovisual recognition of positive and negative affect [151], while other studies report on classification of audiovisual input data into the quadrants in evaluation-activation space [17], [53], [74]. The studies reported in [17], [53], [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 audio-visual 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 has the same coarse emotional states (positive and negative), and then directly used FACS-based labels of facial expressions as audio-visual expression labels.

The data fusion strategies utilized in the current studies on audiovisual affect recognition are either feature-level or decision-level or model-level fusion. Typical examples of feature-level fusion are those reported in [16], [118], [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 a 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], [108]). In the decision-level data fusion, the input coming from each modality is modeled independently and these single-modal recognition results are combined at the end. Since humans display audio and visual expressions in a complementary and 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 that aim at making use of the correlation between audio and visual data streams, and relax the requirement of synchronization of these streams (e.g., [17], [53], [122], [124], [150], [154]). Zeng et al. [154] presented Multi-stream Fused HMM to build an optimal connection among multiple streams from audio and visual channels according to 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 tripled HMM to model 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 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 Neural Networks (NN). Sebe et al. [122] used 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, the research on audiovisual affect recognition has witnessed significant progress in the past few years as follows:

- Efforts have been reported to detect and interpret non-basic genuine (spontaneous) affective displays in terms of coarse affective states such as positive and negative affective states (e.g., [151]), quadrants in evaluation-activation space (e.g., [17], [53], [74]), and application-dependent states (e.g., [122], [154], [97], [108]).
- Few studies have been reported on efforts to integrate other affective cues besides the face and the prosody, such as body and lexical features (e.g., [53], [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], [150]), NN-based fusion (e.g., [53], [74]), and BN-based fusion (e.g., [122]).

4 CHALLENGES

The studies reviewed in the previous section indicate two new trends in the research on automatic human affect recognition: analysis of spontaneous affective behavior and multimodal analysis of human affective behavior including audiovisual analysis, combined linguistic and nonlinguistic analysis, and multi-cue 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 datasets of training and testing material. 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) as well as 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 clips from comedies to subjects, 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 the 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 collec-

tion 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 that could identify the affective state displayed in a video sample and the context in which this affective state was displayed. There are several related issues.

First, it is not clear which kind of metadata need 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 recognition of affective states. However, while this solves the problem of attaining objective facial behavior coding, how to 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 to label 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 to design 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 inter-observer reliability is high, the reliability of the coding is assured. Inter-observer 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 multi-label 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 one hour to manually score 100 still images or a minute of video sequence in terms of AUs [43]. A remedy could be the semi-supervised active learning method [159] that is to combine semi-supervised learning [24] and active learning [51]. The semi-supervised 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 the most ambiguous samples according to the current emotion classifier. More specifically, several promising prototype systems were reported in the past few years that can recognize deliberately produced AUs in either (near-) frontal view face images (e.g., [98], [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 behaviour 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 collection of databases of spontaneous human affective behavior, many of these datasets 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 lack of appropriate agreement of subjects. More specifically, spontaneous displays of emotions, especially in multimedia format, reveal personal and intimate experience; privacy issues jeopardize the public accessibility of many databases.

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

Existing methods for machine analysis of facial affect typically assume that the input data are near frontal- or profile-view face image sequences showing non-occluded facial displays captured under constant lighting condition against a static background. In real interaction environment, such assumption is often invalid. 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 model (e.g., [20], [148]) or multi-view 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], [113]). For instance, it has been shown that spontaneous smiles are longer in 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 onset and offset time 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 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], [158]). However, just few recent studies analyze explicitly the temporal structure of facial expressions (e.g., [98], [99], [135], [132], [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 few studies (e.g., [27], [61]), the existing efforts toward 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 in cognitive science reports that human judgments of behavioral cues are the most accurate when both of the face and the body are taken into account (e.g., [1], [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 automatic coding of human behavior dynamics with respect to both temporal segments (onset, apex, offset, neutral) of various visual cues and temporal correlation between different visual cues (facial, head, and shoulder movements) is that by 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 to 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 to 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 as well as 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 automatic speech recognition (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], [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. dialogue act, repetitions, corrections, and syntactic information) is an even more challenging problem which remains open 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 non-linguistic vocalizations like sighs and yawns [113]. Few efforts towards automatic recognition of non-linguistic vocalizations like laughers [133], [108], cries [97], and coughs [91] have been also recently reported. However, no effort towards 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, multi-visual-cue data fusion would be highly beneficial for machine analysis of human affect, it remains unclear how this should be accomplished. Studies in neurology on 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 to 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 at the end. Decision-level fusion, also called classifier fusion, is now an active area in machine learning and pattern recognition field. Many studies have demonstrated the advantage of classifier fusion over the individual classifiers due to the uncorrelated errors from different classifiers (e.g., [78], [112]). Various classifier fusion methods (fixed rules and trained combiners) have been proposed in 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 at the end as is the case in decision-level fusion.

Model-level fusion or hybrid fusion that aims at com-

bining 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 to model multimodal fusion based on multi-label 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, as well as inclusion of suitable estimations of reliability of each stream. In addition, how to 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 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 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 to make use of information about the context (environment, observed subject, his or her current task) in which the observed affective behavior was displayed. Affects are intimately related to a situation being experienced or imagined by human. Without context, human may misunderstand the observed person's emotion expressions. Yet, with the exception a few studies investigated the influence of context on affect recognition (e.g., [50], [52], [69], [72], [86], [104]), virtually all existing approaches to machine analysis of human affect are context insensitive. Building a context model that includes person ID, gender, age, conversation topic, and workload need the 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 existing methods are tested just on pre-segmented emotion sequences or images, except few studies (e.g., [11], [25]) that use heuristic methods to segment the emotions from videos. Automatic continuous emotion recognition is a dynamic searching process that is to continuously make 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 emotion 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, vocal expressions) have difference 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: Existing methods for 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 dataset. 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

The research in 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 from 2003 [102] and Cowie et al. in 2001 [31]. At that time, a few small-sized datasets of affective displays existed, and almost all methods for machine analysis of human affect were uni-modal, 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 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 datasets 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 machine analysis of human affect as well as on summarizing the issues that have not received sufficient attention but are crucial for advancing machine interpretation of human behavior in naturalistic contexts. The most important of these issues yet to be addressed in the field include the following:

- 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 machine analysis of human affective behavior. Define the appropriate evaluation procedures.
- Develop methods for spontaneous affective behavior analysis that are robust to observed person's arbitrary movement, occlusion, complex and noisy background.
- Devise models and methods for human affect analysis that take into the consideration temporal structures of the modalities and temporal correlations between the modalities (and/or multiple cues), and context (subject, his or her task, environment).
- Develop 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 towards 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 machine analysis of human affect can aid in the creation of a new paradigm for HCI (affect-

sensitive interfaces, socially intelligent environments), and advance the research in several related fields including psychology, psychiatry, and education.

ACKNOWLEDGMENT

This paper is 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. We would like to thank Qiang Ji and anonymous reviewers for encouragement and valuable comments. This work was supported in part by Beckman Postdoctoral Fellowship and NSF CCF 04-26627.

REFERENCES

[1] Ambady, N., Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychological Bulletin, Vol. 111, No. 2, 256-274

[2] Anderson K and McOwan P W (2006). A real-time automated system for recognition of human facial expressions. IEEE Trans. Systems, Man, and Cybernetics- Part B, Vol. 36, No. 1, 96-105

[3] Ang J, Dhillon R, Krupski A, et al. (2002). Prosody-based automatic detection of annoyance and frustration in human-computer dialog, ICSLP.

[4] Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Ambadar, Z., Prkachin, K., Solomon, P. and Theobald, B.J. (2007). The painful face: pain expression recognition using active appearance models. Int'l Conf. Multimodal Interfaces, 9-14

[5] Athanaselis T, Bakamidis S, Dologlou I, Cowie R, Douglas-Cowie E, Cox C (2005). ASR for emotional speech: Clarifying the issues and enhancing performance. Neural Networks, 18:437-444

[6] Austermann, A. Esau, N. Kleinjohann, L. Kleinjohann, B. (2005). Prosody based emotion recognition for MEXI. Int. Conf. Intelligent Robots and Systems, 1138-1144

[7] Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., Kollias, S. (2005). Emotion Analysis in Man-Machine Interaction Systems. Lecture Notes in Computer Science, vol. 3361, 318-328

[8] Banse, R., Scherer, K.R. (1996). Acoustic profiles in Vocal emotion expression. Journal Personality Social Psychology, Vol. 70, No. 3, 614-636

[9] Bartlett M S, Littlewort G, Braathen P, Sejnowski T J and Movellan J R (2003). A prototype for automatic recognition of spontaneous facial actions. Advances in Neural Information Processing Systems, Vol. 15, 1271-1278

[10] Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., and Movellan, J.(2005), Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior, IEEE International Conference on Computer Vision and Pattern Recognition, 568-573

[11] Bartlett M S, Littlewort G, Frank MG, Lainscsek C, Fasel I and Movellan J (2006). Fully automatic facial action recognition in spontaneous behavior. Int. Conf. on Automatic Face and Gesture Recognition, 223-230

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

[13] Batliner A, Hacker C, Steidl S, Noth E, D'Arcy S, et al. (2004). You stupid tin box—Children interacting with the AIBO robot: a cross-linguistic emotional speech. Proceedings LREC.

[14] Blouin, C., and Maffiolo, V. (2005), "A study on the automatic detection and characterization of emotion in a voice service context", Interspeech, Lisbon, 469-472.

[15] Burger S, MacLaren V and Yu H (2002). The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style. In *Proceedings ICSLP*, Denver CO, USA.

[16] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M. et al. (2004), Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information. Int. Conf. Multimodal Interfaces. 205-211

[17] Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaiou, A. and Karpouzis, K.. (2006). Modeling Naturalistic Affective States via Facial and Vocal Expression Recognition. Int. Conf. on Multimodal Interfaces. 146-154

[18] Chang Y, Hu C, Turk, M (2004). Probabilistic expression analysis on manifolds. Proc. Computer Vision and Pattern Recognition, 2:520-527

[19] Chang Y, Hu C, Feris R and Turk M (2006). Manifold based analysis of facial expression. J. Image and Vision Computing, Vol. 24. No.6, 605-614

[20] Chang Y, Vieira M, Turk M, and Velho L (2005). Automatic 3D facial expression analysis in videos. Analysis and Modelling of Faces and Gestures, Proceedings. 3723, pp. 293-307.

[21] Chen, L.S (2000), Joint Processing of Audio-Visual Informa-tion for the Recognition of Emotional Expressions in Human-Computer Interaction, PhD thesis, UIUC

[22] Chen, L, Huang, T. S., Miyasato, T., and Nakatsu, R. (1998). Multimodal human emotion/expression recognition. Int. Conf. on Automatic Face and Gesture Recognition. 396-401

[23] Coan, J.A., Allen, J.J.B. (2007). Handbook of Emotion Elicitation and Assessment. Oxford University Press, New York, USA

[24] Cohen, I., Cozman, F., Sebe, N., Cirelo, M., and Huang, T. S. (2004). Semi-supervised learning of classifiers: theory, algorithms, and their applications to human-computer interaction. IEEE Trans on Pattern Analysis and Machine Intelligence, Vol. 26, No. 12, 1553-1567

[25] Cohen, L., Sebe, N., Garg, A., Chen, L., and Huang, T. (2003), Facial expression recognition from video sequences: Temporal and static modeling, Computer Vision and Image Understanding, 91(1-2):160-187

[26] Cohn, J.F. (2006), Foundations of Human Computing: Facial Expression and Emotion, Int. Conf. on Multimodal Interfaces, 233-238

[27] Cohn JF, Reed LI, Ambadar Z, Xiao J, and Moriyama T. (2004). Automatic Analysis and recognition of brow actions and head motion in spontaneous facial behavior. Int. Conf. on Systems, Man & Cybernetics, 1, 610-616

[28] Cohn, J.F. and Schmidt, K.L.(2004). The timing of Facial Motion in Posed and Spontaneous Smiles, International Journal of Wavelets, Multiresolution and Information Processing, 2, 1-12

[29] Cohn JF and Tronick EZ. (1988). Mother Infant Interaction: the sequence of dyadic states at three, six and nine months. Development Psychology, 23, 68-77

[30] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'Feeltrace': an instrument for recording perceived emotion in real time. Proceedings of the ISCA Workshop on Speech and Emotion, 19–24

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

[32] Cowie R, Douglas-Cowie E and Cox C (2005). Beyond emotion archetypes: databases for emotion modeling using neural networks. Neural Networks, 18: 371-388

[33] Dellaert, F., Polzin, T. and Waibel, A. (1996). Recognizing emotion in speech. Int. Conf. on Spoken Language Processing, 1970-1973

[34] Devillers L and Vasilescu I (2004). Reliability of lexical and prosodic cues in two real-life spoken dialog corpora. Proceedings LREC.

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

[36] Vasilescu, I. and Devillers, L (2005). Detection of real-life emotions in call centers. Interspeech.

[37] Devillers L, Vidrascu L, and Lamel L (2005). Challenges in real-life emotion annotation and machine learning based detection. Neural Networks, 18: 407-422

[38] Douglas-Cowie E., Campbell N, Cowie R and Roach P (2003). Emotional Speech: towards a new generation of database. Speech Communication, 40(1-2): 33-60

[39] Duric, Z., Gray, W.D., Heishman, R., Li, F., Rosenfeld, A., Schoelles, M.J., Schunn, C., Wechsler, H. (2002). Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. Proceedings of the IEEE, Vol. 90, No. 7, 1272-1289

[40] Ekman P. (1972). Universals and cultural differences in facial expressions of emotion. Nebr. Symp. Motiv. 1971, 207-283

[41] Ekman, P., editor (1982). Emotion in the human face. Cambridge University Press, New York, 2nd edition

[42] Ekman, P. (1994), Strong Evidence for Universals in Facial Expressions: A Reply to Russell's Mistaken Critique, Psychological Bulletin, 115(2): 268-287

[43] Ekman, P., Friesen, W.V., Hager, J.C. (2002). Facial Action Coding System. A Human Face, Salt Lake City, USA

[44] Ekman, P., Huang, T.S., Sejnowski, T.J. and Hager, J.C., (Eds.), (1993). NSF Understanding the Face, A Human Face eStore, Salt Lake City, USA, (see Library).

[45] Ekman P, Matsumoto D, and Friesen WV. (2005). Facial Expression in Affective Disorders. In What the Face Reveals. Edited by Ekman P and Rosenberg EL. 429-439

[46] Ekman P. and Oster H. (1979). Facial expressions of emotion. Ann. Rev. Psychol. 1979, 30:527-554

[47] Ekman P. and Rosenberg E.L. (2005). What the face reveals: basic and applied studies of spontaneous expression using the facial action coding system. 2nd edition, Oxford University Press.

[48] El Kaliouby R and Robinson P (2004). Real-time Inference of complex mental states from facial expression and head gestures. Computer Vision and Pattern Recognition Workshop, Vol. 3, 154

[49] Fasel, B. and Luttin, J. (2003). Automatic facial expression analysis: Survey. Pattern Recognition, 36(1): 259-275

[50] Fasel B, Monay F and Gatica-Perez D (2004). Latent semantic analysis of facial action codes for automatic facial expression recognition. ACM Int. Workshop on Multimedia Information Retrieval, 181-188

[51] Fiechter, C-N. (1994). Efficient reinforcement learning. ACM Conf. on Computational Learning Theory, 88-97

[52] Forbes-Riley K and Litman D (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. Proc. Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)

[53] Fragopanagos, F. and Taylor, J.G. (2005), Emotion recognition in human-computer interaction, Neural Networks, 18: 389-405

[54] Fried E. (1976). The impact of nonverbal communication of facial affect on children's learning. PhD thesis, Rutgers University, New Brunswick, NJ

[55] Furnas, G., Landauer, T., Gomes, L., and Dumais, S. (1987). The vocabulary problem in human-system communication, Communications of the ACM, Vol. 30, No. 11, 964-972.

[56] Go HJ, Kwak KC, Lee DJ, and Chun MG. (2003). Emotion recognition from facial image and speech signal. Int. Conf. of the Society of Instrument and Control Engineers. 2890-2895

[57] Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., and Kajarekar, S. (2006). Combining prosodic, lexical and cepstral systems for deceptive speech detection. Proc. Int. Conf. on Acoustics, Speech and Signal Processing, I:1033-1036

[58] Greenwald M, Cook E and Lang P. (1989). Affective judgment and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. J. Psychophysiol. 3:51-64

[59] Gross, R. (2005) Face databases. In: Handbook of Face Recognition, Li S Z., Jain A.K., (Eds.), Springer, New York, USA, 301-328

[60] Gu H, Ji Q (2004). An Automated Face Reader for Fatigue Detection. Int. Conf. Automatic Face and Gesture Recognition. 111-116

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

[62] Gunes, H. and Piccardi, M. (2005). Fusing Face and Body Display for Bi-Modal Emotion Recognition: Single Frame Analysis and Multi-Frame Post Integration. Int. Conf. on Affective Computing and Intelligent Interaction, 102 – 111

[63] Gunes, H. and Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. Int. Conf. on Pattern Recognition, Vol. 1, 1148-1153

[64] Guo G and Dyer C R (2005). Learning from examples in the small sample case – face expression recognition. IEEE Trans. Systems, Man and Cybernetics – Part B, Vol.35, No.3, 477-488

[65] Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S. (2005). Distinguishing Deceptive from Non-Deceptive Speech. Interspeech, 1833-1836

[66] Hoch, S., Althoff, F., McGlaun, G., Rigoll, G. (2005), Bimodal fusion of emotional data in an automotive environment, ICASSP, Vol. II, 1085-1088, 2005

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

[68] Jaimes, A. and Sebe, N. (2005). Multimodal human computer interaction: a survey. Workshop on Human Computer Interaction in conjunction with ICCV.

[69] Ji Q, Lan P and Looney C (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. IEEE SMC-Part A, Vol. 36, No.5, 862-875

[70] Juslin, P.N., Scherer, K.R. (2005). Vocal expression of affect. In The New Handbook of Methods in Nonverbal Behavior Research. Harrigan, J., Rosenthal, R., Scherer, K., Eds. Oxford University Press, Oxford, UK

[71] Kanade, T., Cohn, J., and Tian, Y. (2000), Comprehensive Database for Facial Expression Analysis, In Proceeding of International Conference on Face and Gesture Recognition, 46-53

[72] Kapoor, A., Burleson, W., and Picard, R. W. (2007), Automatic prediction of frustration. Int. Journal of Human-Computer Studies. Vol. 65(8), 724-736.

[73] Kapoor, A. and Picard, R. W. (2005). Multimodal affect recognition in learning environment. ACM Int'l Conf. on Multimedia, 677-682

[74] Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., and Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal, and bodily expression recognition, Lecture Notes in Artificial Intelligence, vol. 4451, 91-112.

[75] Keltner D (1995). Signs of appeasement: evidence for the distinct displays of embarrassment, amusement and shame. Journal of Personality and Social Psychology, 68(3). 441-454

[76] Kobayashi, H., and Hara, F. (1991). The recognition of basic facial expressions by neural network. Proc. Int'l Joint Conf. on Neural Networks, 460-466.

[77] Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features ans support vector machine. IEEE Trans. On Image Processing, 16(1): 172-187

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

[79] Kwon, O.W., Chan, K., Hao, J., Lee, T.W (2003), Emotion Recognition by Speech Signals, EUROSPEECH.

[80] Laskowski, K. and Burger, S. (2006). Annotation and Analysis of Emotionally Relevant Behavior in the ISL Meeting Corpus, LREC, Genoa, Italy.

[81] Lee, C. and Elgammal, A. (2005). Facial expression analysis using nonlinear decomposable generative models. In IEEE International Workshop on Analysis and Modeling of Faces and Gestures

[82] Lee C and Narayanan (2003). Emotion recognition using a data-driven fuzzy inference system. In Proc. Eurospeech, 157-160

[83] Lee C M Narayanan, S.S. (2005). Toward detecting emotions in spoken dialogs. IEEE Tran. Speech and Audio Processing, Vol. 13(2): 293-303

[84] Liscombe, J., Hirschberg, J., Venditti, J.J. (2005). Detecting Certainness in Spoken Tutorial Dialogues. Interspeech.

[85] Lisetti, C.L., Nasoz, F. (2002). MAUI: A multimodal affective user interface. Proc. Int'l Conf. Multimedia, 161-170

[86] Litman, D.J. and Forbes-Riley, K. (2004), Predicting Student Emotions in Computer-Human Tutoring Dialogues. In Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL), July

[87] Littlewort, G.C., Bartlett, M.S. and Lee, K. (2007). Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. Int'l Conf. Multimodal Interfaces, 15-21

[88] Lucey, S., Ashraf, A.B., and Cohn, J.F. (2007), Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face. In Face Recognition, Delac, K. and Grgic, M., Eds. Vienna, Austria: I-Tech Education and Publishing, 275-286

[89] Maat, L., and Pantic, M. (2006). Gaze-X: Adaptive affective multimodal interface for single-user office scenarios, Proc. ACM Int'l Conf. Multimodal Interfaces, 171-178

[90] Mase, K. (1991). Recognition of facial expression from optical flow. IEICE Trans. E74(10). 3474-3483

[91] Matos, S., Birring, S.S., Pavord, I.D. and Evans, D.H. (2006). Detection of cough signals in continuous audio recordings using HMM. IEEE Trans. Biomedical Engineering, Vol. 53, No. 6, 1078-1083.

[92] Mehrabian, A. (1968). Communication with words. Psychology Today, 2(4): 53-56

[93] Mozziconacci, S. (2002). Prosody and Emotions. Proc. Speech Prosody Aix-en-Provence, 1-9.

[94] Neiberg D, Elenius K, and Laskowski K. (2006). Emotion Recognition in Spontaneous Speech Using GMM. Int. Conf. on Spoken Language Processing, 809-812

[95] O'Toole A J, Harms J, Snow S L, Hurst D R, Pappas M R, et al. (2006). A Video Database of Moving Faces and People. IEEE PAMI, VOL. 27, NO. 5, MAY 2005, 812-816

[96] Oudeyer, P-Y. (2003). The production and recognition of emotions in speech: features and algorithms. Int'l Journal Human-Computer Studies, 59, 157-183.

[97] Pal P, Iyer A N and Yantorno R E (2006). Emotion detection from infant facial expressions and cries. In Proc. Int'l Conf. Acoustics, Speech & Signal Processing, 2, pp. 721-724, 2006.

[98] Pantic, M., and Bartlett, M.S. (2007). Machine analysis of facial expressions. In Face Recognition, Delac, K. and Grgic, M., Eds. Vienna, Austria: I-Tech Education and Publishing, 377-416

[99] Pantic, M., and Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments form face profile imgae sequences. IEEE Trans. Systems, Man and Cybernetics – Part B, Vol. 36, No.2, 433-449

[100] Pantic, M., Pentland, A., Nijholt, A., and Huang, T.S. (2006), Human Computing and Machine Understanding of Human Behavior: A Survey, Int. Conf. on Multimodal Interfaces, 239-248

[101] Pantic, M., and Rothkrantz, L.J.M. (2000). Automatic analysis of facial expressions—the state of the art. IEEE PAMI, Vol.22, No.12, 1424-1445

[102] Pantic M., and Rothkrantz, L.J.M. (2003), Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE, Vol. 91, No. 9, Sept., 1370-1390

[103] Pantic, M., and Rothkrantz, L.J.M. (2004). Facial action recognition for facial expression analysis from static face images. IEEE Trans. On Systems, Man and Cybernetics-Part B, Vol. 34, No. 3, 1449-1461

[104] Pantic, M., and Rothkrantz, L.J.M. (2004). Case-based reasoning for user-profiled recognition of emotions from face images. Int. Conf. Multimedia & Expo, 391-394

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

[106] Pantic, M., Valstar, M.F, Rademaker, R. and Maat, L. (2005), Web-based database for facial expression analysis, Int. Conf. on Multimedia and Expo, 317–321

[107] Pentland, A. (2005). Socially aware, computation and communication, IEEE Computer, Vol.38, 33-40

[108] Petridis, S. and Pantic, M. (2008). Audiovisual discrimination between laughter and speech. Int'l Conf. Acoustics, Speech, and Signal Processing.

[109] Picard, R.W. (1997). Affective Computing, MIT Press, Cambridge.

[110] Plutchik R. (1980). Emotion: A psychoevolutionary synthesis. New York: Harper and Row.

[111] Roisman, G.I., Tsai, J.L., Chiang, K.S.(2004). 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, 776-789

[112] Roli, F., Kittler, J., et al.., eds., (2001-2005). Int. Workshop Multiple Classifier Systems (MCS).

[113] Russell J.A., Bachorowski J. and Fernandez-Dols J. (2003). Facial and vocal expressions of emotion. Ann. Rev. Psychol. 54:329-349

[114] Russell J and Mehrabian A. (1977). Evidence for a three-factor theory of emotions. J. Res. Personality, 11: 273-294

[115] Samal A and Iyengar P A (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. Pattern Recognition, Vol. 25, No.1, 65-77

[116] Sander D. Grandjean D. and Scherer K.R. (2005). A system approach to appraisal mechanisms in emotion. Neural Networks. 18: 317-352

[117] Scherer K.R. (1999). Appraisal theory. In Dalgleish T and Power M J (Eds.), Handbook of cognition and emotion, New York: Wiley, 637-663

[118] Schuller, B., Muller, R., Hornler, B., Hothker, A., Konosu, H. and Rigoll, G. (2007). Audiovisual recognition of spontaneous Interest within conversations. Int. Conf. on Multimodal Interfaces, 30-37

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

[120] Schuller, B., Villar, R. J., Rigoll, G., Lang, M. (2005). Meta-Classifiers in acoustic and linguistic feature fusion-based affect recognition. Int. Conf. on Acoustics, Speech, and Signal Processing, 325-328

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

[122] Sebe, N., Cohen, I., Gevers, T. and Huang, T.S. (2006). Emotion recognition based on joint visual and audio cues. Int. Conf. on Pattern Recognition, 1136-1139

[123] Sebe, N., Lew, M.S., Cohen, I., Sun, Y., Gevers, T., Huang, T.S.(2004), Authentic Facial Expression Analysis, Int. Conf. on Automatic Face and Gesture Recognition

[124] Song, M., Bu, J., Chen, C., and Li, N. (2004), Audio-visual based emotion recognition—A new approach, Int. Conf. Computer Vision and Pattern Recognition. 2004, 1020-1025

[125] Steidl, S., Levit, M., Batliner, A, Noth, E., and Niemann, H. (2005), "Off all things the measure is man" Automatic classification of emotions and inter-labeler consistency, ICASSP, vol.1, 317-320

[126] Stein, B., Meredith, M.A. (1993). The Merging of Senses. MIT Press, Cambridge, USA

[127] Suwa, M.,Sugie, N., Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. Int. Joint Conference on Pattern Recognition. 408-410

[128] Tao, H. and Huang, T.S. (1999), Explanation-based facial motion tracking using a piecewise Bezier volume deformation mode, IEEE CVPR, vol.1, pp. 611-617,

[129] Teeters, A., Kaliouby, R. E., and Picard, R. W. (2006). Self-Cam: Feedback From What Would Be Your Social Partner. ACM SIGGRAPH, Research Posters, p. 138

[130] Tian Y L, Kanade T and Cohn J F (2005). Facial expression analysis. In: Handbook of Face Recognition, Li S Z and Jain A K (Eds.), Springer, New York, USA, 247-276

[131] Tomkins SS. (1962). Affect, Imagery, Consciousness, Vol. 1. New York: Springer

[132] Tong, Y., Liao, W. and Ji, Q. (2007). Facial action unit recognition by exploiting their dynamics and semantic relationships. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 10, 1683-1699

[133] Truong K P and van Leeuwen D A (2007). Automatic discrimination between laughter and speech. Speech Communication, 49: 144-158.

[134] Valstar, M.F., Gunes, H. and Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. Int'l Conf. Multimodal Interfaces, 38-45.

[135] Valstar, M., Pantic, M., Ambadar, Z., and Cohn, J.F. (2006). Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. Int. Conf. on Multimedia Interfaces. 162-170

[136] Valstar, M., Pantic, M., and Patras, I. (2004). Motion history for facial action detection from face video. Int. Conf. Systems, Man and Cybernetics, Vol.1, 635-640

[137] Wang, H. and Ahuja, N. (2003). Facial expression decomposition. IEEE International Conference on Computer Vision, p.958

[138] Wang, Y. and Guan, L.(2005), Recognizing human emotion from audiovisual information, ICASSP, Vol. II, 1125-1128

[139] Wang, J., Yin, L., Wei, X., and Sun, Y. (2006). 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. IEEE Conference on Computer Vision and Pattern Recognition, 2:1399-1406

[140] Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, Vol. 54, 1063-1070.

[141] Wen, Z. and Huang, T.S. (2003). Capturing Subtle Facial Motions in 3D Face Tracking. Int. Conf. on Computer Vision, 1343-1350

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

[143] Whitehill J. and Omlin, C. W. (2006). Haar features for FACS AU recognition. Int. Conf. on Automatic Face and Gesture Recognition, 217-222

[144] Williams, A. C. de C. (2002) Facial expression of pain: An evolutionary account. Behavioral and Brain Sciences, Vol. 25, No. 4, 439–488

[145] Williams, C. and Stevens, K. (1972). Emotions and speech: Some acoustic correlates. Journal of the Acoustic Society of America, 52(4), 1238-1250

[146] Xiao J, Moriyama T, Kanade T and Cohn J F (2003). Robust full-motion recovery of head by dynamic templates and re-registration techniques. Int. J. Imaging Systems and Technology, Vol. 13, No.1, 85-94

[147] Yeasin M., Bullot B. and Sharma R. (2006), Recognition of facial expressions and measurement of levels of interest from video, IEEE Trans. On Multimedia, Vol.8, No. 3, June, 500-507

[148] Yin L, Wei X , Sun Y, Wang J, Rosato M J (2006). A 3D facial expression database for facial behavior research. Int. Conf. on Automatic Face and Gesture Recognition, 211- 216

[149] Zeng, Z., Fu, Y., Roisman, G.I., Wen, Z., Hu, Y., and Huang, T.S. (2006). Spontaneous Emotional Facial Expression Detection. Journal of Multimedia, 1(5): 1-8.

[150] Zeng, Z., Hu, Y., Liu, M., Fu, Y. and Huang, T.S.(2006), Training Combination Strategy of Multi-stream Fused Hidden Markov Model for Audio-visual Affect Recognition, in Proc. ACM Int'l Conf. on Multimedia, 2006, 65-68

[151] Zeng, Z., Hu, Y., Roisman, G.I., Wen, Z., Fu, Y., and Huang, T.S. (2007), Audio-visual Spontaneous Emotion Recognition, In Artificial Intelligence for Human Computing, Eds: Huang T.S., Nijholt A., Pantic M., and Pentland A., Springer, 72-90.

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

[153] Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T.S., Roth, D., and Levinson, S. (2004), Bimodal HCI-related Emotion Recognition, Int. Conf. on Multimodal Interfaces, 137-143.

[154] Zeng, Z., Tu, J., Pianfetti , P., Liu, M., Zhang, T., Zhang Z., Huang T S and Levinson S (2005), Audio-visual Affect Recognition through Multi-stream Fused HMM for HCI, Int. Conf. Computer Vision and Pattern Recognition. 967-972

[155] Zeng, Z., Tu, J., Liu, M., Huang, T.S., Pianfetti, B., Roth D. and Levinson, S. (2007), Audio-visual Affect Recognition, IEEE Transactions on Multimedia, Vol. 9, No. 2, February, 424-428

[156] Zeng, Z., Zhang, Z., Pianfetti, B., Tu, J., and Huang, T.S. (2005), Audio-visual Affect Recognition in Activation-evaluation Space, Int. Conf. on Multimedia & Expo, 828-831.

[157] Zhang T, Hasegawa-Johnson M and Levinson S E (2004). Children's Emotion Recognition in an Intelligent Tutoring Scenario, Interspeech 2004.

[158] Zhang Y and Ji Q (2005). Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequences. IEEE Trans. Pattern Anal. Mach. Intell. 27(5): 699-714

[159] Zhou, Z.-H., Chen, K.-J. and Dai, H.-B., Enhancing relevance feedback in image retrieval using unlabeled data. ACM Transactions on Information Systems, 2006, 24(2): 219-244

[160] Zhu, Z. and Ji, Q. (2006). Robust Real-Time Face Pose and Facial Expression Recovery. IEEE Conference on Computer Vision and Pattern Recognition, 1: 681-688

Zhihong Zeng received his PhD in Institute of Automation, Chinese Academy of Sciences in 2002. He is currently Beckman Postdoctoral Fellow at Beckman Institute, UIUC. His research interests include multimodal affective computing, multimodal human computer interaction and computer vision. He is an IEEE member.

Maja Pantic received the MSc and PhD degrees in Computer Science from Delft University of Technology, The Netherlands, in 1997 and 2001. She is Reader in Multimodal HCI at Imperial College London, Computing Department, and Professor in Affective and Behavioural Computing at the University of Twente, Computer Science Department. 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 an IEEE senior member. She is an Associate Editor of IEEE Trans. on Systems, Man and Cybernetics - Part B, and of Image and Vision Computing Journal. She is a guest editor, organizer and committee member of over 10 major journals and conferences.

Glenn I. Roisman received his PhD from the University of Minnesota in 2002. He is currently assistant professor in the Department of Psychology at UIUC. His research interests concern social and emotional development across the lifespan. He has published over twenty-five scholarly journal articles and chapters, and received the Society for Research in Child Development's Award for Early Research Contributions in 2007.

Thomas S. Huang received his Sc.D. from MIT in 1963. He is William L. Everitt Distinguished Professor of Electrical and Computer Engineering, and Co-chair of Human Computer Intelligent Interaction Initiative (HCII) in the Beckman Institute, UIUC. His professional interests are computer vision, image compression and enhancement, pattern recognition, and multimodal signal processing. He has more than 80 honors, awards and outstanding achievements, including Member of National Academy of Engineering; Fellow of IEEE; Foreign Member of Chinese Academy of Sciences.