Chapter 14
Emotion Recognition Based on Multimodal Information

Zhihong Zeng, Maja Pantic, and Thomas S. Huang

14.1 Introduction

Here is a conversation between an interviewer and a subject occurring in an Adult Attachment Interview (Roisman, Tsai, & Chiang, 2004). AUs are facial action units defined in Ekman, Friesen, and Hager (2002).

The interviewer asked: "Now, let you choose five adjective words to describe your childhood relationship with your mother when you were about five years old, or as far back as you remember."

The subject kept smiling (lip corner raiser AU12) when listening. After the interviewer finished the question, the subject looked around and lowered down her head (AU 54) and eyes (AU 64). Then she lowered and drew together the eyebrows (AU4) so that severe vertical wrinkles and skin bunching between the eyebrows appeared. Then her left lip raise[d] (Left AU10), and finger scratched chin.

After about 50 second silence, the subject raise her head (AU53) *and brow (AU1+AU2), and asked with a smile (AU12): "Should I . . . give what I have now?"*

The interviewer response with smiling (AU12): "I guess, those will be when you were five years old. Can you remember?"

The subject answered with finger touching chin: "Yeap. Ok. Happy (smile, AU 6+AU12), content, dependent, (silence, then lower her voice) what is next (silent, AU4+left AU 10), honest, (silent, AU 4), innocent."

Z. Zeng (✉)
University of Illinois at Urbana-Champaign, Imperial College London
e-mail: zhzeng@ifp.uiuc.edu

M. Pantic
Imperial College London / University of Twente, Netherlands
e-mail: m.pantic@imperial.ac.uk

T.S. Huang
University of Illinois at Urbana-Champaign, Imperial College London
e-mail: huang@ifp.uiuc.edu

J.H. Tao, T.N. Tan (eds.), *Affective Information Processing,*

This is an exemplar interaction occurring in a natural human interaction setting where the verbal and nonverbal behavior involves the coordination of multiple modalities (facial expression, speech (linguistic and paralinguistic information), gesture, gaze, head movement, and context). Generally, humans consciously or unconsciously use these modalities to express and interpret emotional behavior in such a way that the interaction can go smoothly. Each of these modalities has a unique contribution in the exchange of information of human behavior, as described in the other chapters. However, unimodal analysis is not sensitive enough to capture all the emotion content of the interactions, nor reliable enough to understand the meaning of the emotion behavior.

Take the above conversation as an instance. If we just listen to what the participants said, we will miss much behavior during the silence. On the other hand, if we just watch the facial actions without the audio channel, we may not reliably interpret the smiles (AU12, AU6+AU12), frowns (AU4), head raise (AU53), brow raise (AU2), and unilateral upper lip raise (left AU10). The subject's complex behavior (e.g., greeting, thinking, uncertainty, and induced emotion) must be understood by integrating physical features from multiple modalities (facial expression, speech, gesture, gaze, and head movement) with context. For example, the first smile of the subject may just mean that she was following what the interviewer was saying whereas the last smile of the subject may mean joy induced by her childhood experience recall.

Many psychological studies have theoretically and empirically demonstrated the importance of integration of information from multiple modalities to yield a coherent representation and inference of emotions (e.g., Ambady & Rosenthal, 1992; Scherer, 1999; Russell, Bachorowski, & Fernandez-Dols, 2003; Yoshimoto, Shapiro, O'Brian, & Gottman, 2005). Emotion research requires the study of the configuration of these emotion-related modalities, and multimodal integration is a key to understanding how humans efficiently and reliably express and perceive human emotional behavior.

With the development of science and technology, more and more researchers from the engineering and psychological communities have explored the possibility of using computers to automatically analyze human emotion displays. We have witnessed significant progress in the field of machine analysis of human emotion behavior, especially facial expression and vocal expression (Picard, 1997; Cowie et al.., 2001; Pantic & Rothkrantz, 2003; Pentland, 2005; Cohn, 2006).

It has also been shown by several studies that integrating the information from multiple modalities leads to improvement of machine recognition performance of emotion over unimodal approaches (Song, Bu, Chen, & Li, 2004; Fragopanagos & Taylor, 2005; Caridakis, Malatesta, Kessous, Amir, Paouzaiou, & Karpouzis, 2006; Zeng et al., 2007b; Zeng, Tu, Pianfetti, & Huang, 2008b).

The improved reliability of multimodal approaches can be explained from an engineering perspective. Current techniques for different modalities have different limitations; for instance, detection and tracking of facial expressions are sensitive to head pose, clutter, and variations in lighting conditions, speech processing is sensitive to auditory noise, and detection of physiological responses is influenced by

intrusion of wearable devices. Multimodal fusion provides a possibility to make use of complementary information from multiple modalities and reduce the current technical limitations in automatic emotion analysis.

Based on the fact that change in emotion plays an essential role in our daily life, especially in social interaction, some researchers have explored the possibility of enabling the computer to recognize human emotion behavior, with the goal of building a natural and friendly human–computer interaction (HCI) environment. Examples of affect-sensitive, multimodal HCI systems include the system of Lisetti and Nasoz (2002), which combines facial expression and physiological signals to recognize the user's emotion such as fear and anger and then to adapt an animated interface agent to mirror the user's emotion; the multimodal system of Duric et al., (2002), which applies a model of embodied cognition that can be seen as a detailed mapping between the user's emotional states and the types of interface adaptations; the proactive HCI tool of Maat and Pantic (2006) capable of learning and analyzing the user's context-dependent behavioral patterns from multisensory data and of adapting the interaction accordingly, the automated Learning Companion of Kapoor, Burleson, and Picard (2007) 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 needs help; and the multimodal computer-aided learning system at Beckman Institute UIUC [1] illustrated in Figure 14.1 where the computer avatar offers an appropriate tutoring strategy based on the information of user's

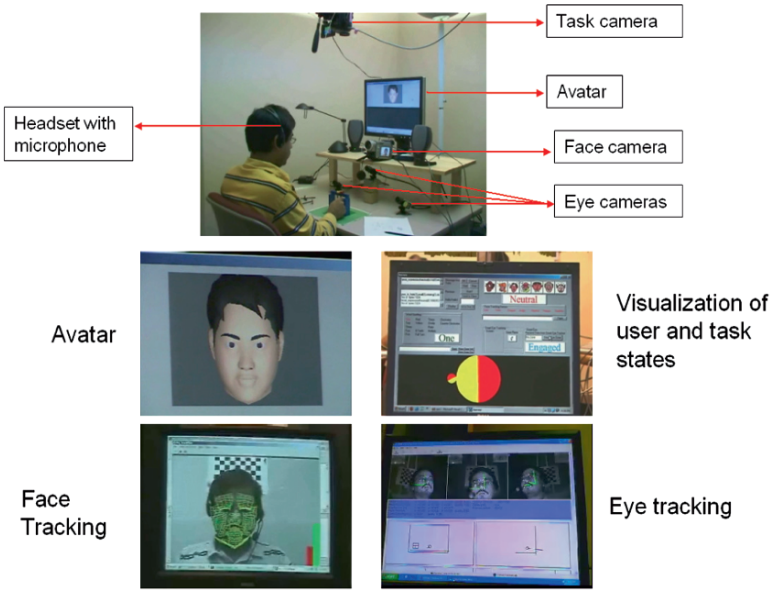

Fig. 14.1 A prototype of multimodal computer-aided learning system.

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

facial expression, eye movement, keywords, eye movement, and task state. These systems represent initial efforts towards the future socially aware, multimodal HCI.

This chapter is intended to provide an overview of the research on automatic human emotion recognition based on the information from multiple modalities (audio, visual, and physiological responses). Multimodal fusion includes audio-visual fusion, visual-physiological fusion, multi-audio-cue fusion (linguistic and paralinguistic information), multi-visual-cue fusion (facial expression, gaze, head movement, and body movement), and audio-visual-physiological fusion. Although the combination of all modalities (i.e., audio-visual-physiological fusion) is expected to be the best choice for emotion analysis (Yoshimoto, Shapiro, O'Brian, & Gottman, 2005), there is no reported effort toward inclusion of all the modalities into an automatic emotion-sensitive computing system.

The chapter is organized as follows. Section 14.2 provides a brief description of emotion, emotion expression, and perception, from a psychological perspective. Section 14.3 provides a detailed review of the related studies, including multimedia emotion databases, existing automatic multimodal emotion recognition methods, and recent authors' efforts toward this research direction. Section 14.4 discusses some of the challenges that researchers face in this field. Finally we draw conclusions from this chapter.

14.2 Human Emotion Expression and Perception

Although the researchers in the emotion-related research communities (psychology, linguistics, neuroscience, anthropology, and other related disciplines) have not reached consensus on the answer to the question, "What is emotion," a widespread accepted description of emotion consists of multiple components (cognitive appraisal, action tendencies, motor expression, physiological symptoms, subjective feeling) that represent the different aspects of emotion (Scherer, 1999; Cohn, 2006). All of these components concurrently work and function relative to each other during an emotion episode.

Unfortunately, the states of cognitive appraisal, action tendencies, and subjective feeling are not observable, so the emotional states can only be inferred through observed emotion expression, detected physiological symptoms, and self-report if possible. Human judgment of emotion is mainly based on emotion expression (facial expression, speech, gesture, body movement, and gaze). If two individuals are close enough or touch each other, they may perceive some physiological symptoms, such as heartbeat and sweat on the skin, which provide an additional perspective to infer emotion reaction.

The emotion inference from observable behavior and physiological symptoms is definitely an ill-posed problem, so the context of emotion induction acts as a constraint to reduce the ambiguity of emotion judgment. The context can be any information to characterize the situation in which an emotion behavior occurs, including who the emotion expresser and receiver are, what the expresser is doing, when and

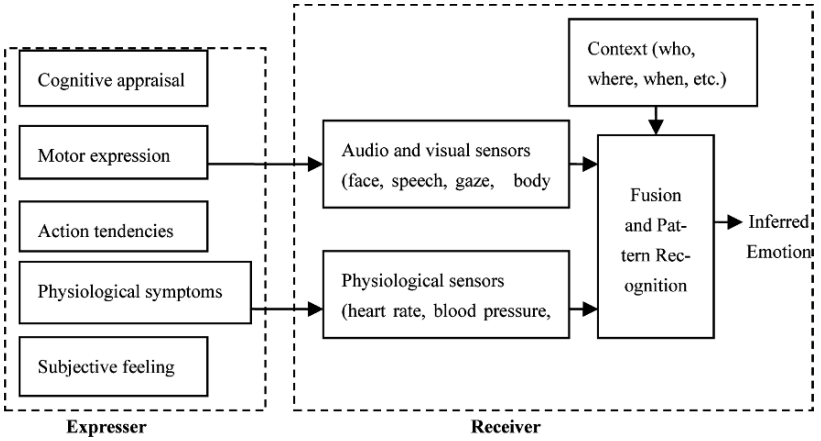

Fig. 14.2 A emotion emission diagram from expresser to receiver.

where the emotion behavior occurs, and so on. Figure 14.2 illustrates a diagram of multimodal emotion expression and perception in which the audio and visual sensors of the receiver (i.e., human eyes and ears) capture the emotion expressions (facial expression, speech, body movement, and gaze), and physiological sensors (i.e., skin) track the physiological responses, and a fusion and pattern recognition module (brain) integrates all related information of audio and visual expressions and physiological responses with the context and makes the judgment of emotion.

In the last three decades, we have witnessed significant progress toward the understanding of emotion behavior and psychological responses, especially based on single modalities of nonverbal behaviors (facial and vocal expressions; Ekman & Friesen, 1975, Ekman, 1982, Ekman & Rosenberg, 2005; Russell, Bachorowski, & Fernandez-Dols, 2003; Harrigan, Rosenthal, & Scherer, 2005). Some promising manual coding systems have been proposed, such as the Facial Action Unit System (Ekman, Friesen, & Hager, 2002) for facial expression and the Feeltrace system (Cowie, Douglas-Cowie, Savvidou, McMahon, Sawey, & Schröder, 2000) for audiovisual expression.

Psychologists have various opinions about the importance of different nonverbal cues in human emotion judgment. Ekman (1982) found that the relative contributions of facial expression, speech, and body gesture to emotion judgment depend both on the emotional state and the environment where the emotional behavior occurs whereas some studies (e.g., Ambady & Rosenthal, 1992) indicated that a facial expression in the visual channel is the most important emotion cue. Many studies have theoretically and empirically demonstrated the advantage of integration of multiple modalities in human emotion perception over single modalities (Ambady & Rosenthal, 1992; Scherer, 1999; Russell et al., 2003; Yoshimoto et al., 2005).

A large number of studies in psychology, linguistics, and neuroscience confirm the correlation between some emotional displays (especially prototypical emotions) and specific audio and visual signals (e.g., Ambady & Rosenthal, 1992;

Cowie et al., 2001; Russell et al., 2003; Ekman & Rosenberg, 2005) and psychological responses (Picard, Vyzas, & Healey, 2001; Stemmler, 2003). 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 emotional behavior rather than posed exaggerated displays.

Recent research results indicated that body movement and gaze significantly facilitate the expression and perception of emotion. Specifically, body movement (i.e., head, limbs, and torso) can provide information of the emotion intensity together with facial and vocal expression, and gaze changes (direct versus averted gaze (Adams & Kleck, 2003), widened eyes, tensed lower lids (Ekman & Friesen, 1975)) are relevant to feeling and attitudes.

Detection of physiological change provides a window into the inner world of human emotion experience (Picard et al., 2001; Stemmler, 2003). But current technology of physiological sensors limits its application in emotion analysis: the available wearable physiological sensors imply wiring the subjects, which make it difficult to elicit emotional displays without influencing results; it can be tricky to gather accurate physiological data because the physiological sensing systems are influenced by other nonemotional factors, including skin–sensor interface influence (positioning of the sensors, application of amounts of gel), human activity, and hormones.

Thus, the progress of physiological research of emotion has been overshadowed by that of audio and visual expression research. Recently, some new physiological sensors have been applied for automatic emotion analysis, such as a sensor mouse in the study (Liao, Zhang, Zhu, Ji, & Gray, 2006), sensor chair in the study (Kapoor & Picard, 2005, Kapoor, Burleson, & Picard, 2007), and wireless noninvasive armband in the study (Lisetti & Nasoz, 2004).

These emotion-related modalities (audio and visual expression, physiological responses) are often studied separately. This precludes finding evidence of the correlation between them. Relatively few studies (Yoshimoto et al., 2005) proposed some pragmatic scheme to integrate all of these modalities to analyze emotion in the interaction setting (such as the conversation between committed couples). However, translating this scheme into one engineering framework for the purposes of automatic emotion recognition remains challenging.

On the other hand, a growing body of research in cognitive sciences argues that the dynamics of human behavior are crucial for its interpretation (Scherer, 1999; Russell et al., 2003; Ekman & Rosenberg, 2005). 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., Cohn et al., 2004; Ekman & Rosenberg, 2005; Valstar, Pantic, Ambadar, & Cohn, 2006) as well as for categorization of complex behaviors like shame, and amusement (e.g., Ekman & Rosenberg, 2005). Based on these findings, we may expect that temporal dynamics of each modality separately and their temporal correlations play an important role in the recognition of human naturalistic emotion behavior. However, these are virtually unexplored areas of research.

Another largely unexplored area of research is the relationship between emotion and context. 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 (Russell et al., 2003).

14.3 Multimodal Emotion Recognition

With the advance of technology, an increased number of studies on machine analysis of multimodal human emotion displays have emerged in recent years. It has been shown by several experimental studies that integrating the information from multiple modalities leads to an improved performance of emotion behavior recognition. The study of Chen, Huang, Miyasato, and Nakatsu in 1998 represents an early attempt toward multimodal emotion recognition, focusing on audiovisual fusion. In this section we first offer a brief overview of the existing databases of multimedia recordings of human emotion displays, which provide the basis of automatic emotion analysis. Next we examine available multimodal computing methods for emotion recognition. We focus here on the efforts recently proposed in the literature that represent multimodal approaches to the problem of multimodal human affect recognition. For exhaustive surveys of the past work in machine analysis of multimodal emotion expressions, readers are referred to the survey papers by Cowie et al.(2001), Pantic and Rothrantz (2003, 2006), Sebe, Cohen, and Huang (2005), and Zeng, Pantic, Roisman, and Huang, (2008a).

14.3.1 Multimodal Databases

Authentic emotion expressions are difficult to collect because they are relatively rare and short-lived, and filled with subtle context-based changes that make it difficult to elicit emotion displays without influencing 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 emotion displays were based on the "artificial" material of deliberately expressed emotions, especially six basic emotions (i.e., happiness, sadness, anger, disgust, anger, surprise), elicited by asking the subjects to perform a series of emotional expressions in front of a camera and/or microphone. As a result, the majority of the existing systems for human emotion recognition aim at classifying the input expression as the basic emotion category (Cowie et al., 2001; Pantic & Rothkrantz, 2003; Sebe et al., 2005).

However, increasing evidence suggests that deliberate behavior differs in visual appearance, audio profile, and timing from spontaneously occurring behavior. For example, Whissell shows that the posed nature of emotions in spoken language may differ in the choice of words and timing from corresponding performances in natural settings (Whissell, 1989). 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 (Ekman & Rosenberg, 2005).

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 (Cohn, Reed, Ambadar, Xiao, & Moriyama, 2004; Ekman & Rosenberg, 2005). 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 (Valstar et al., 2006). It is not surprising, therefore, that methods of automated human emotion analysis that have been trained on deliberate and often exaggerated behaviors usually fail to generalize to the subtlety and complexity of spontaneous emotion behavior.

These findings and the general lack of a comprehensive reference set of audio and/or visual recordings of human emotion displays motivated several efforts aimed at the development of datasets that could be used for training and testing of automatic systems for human emotion analysis. Table 14.1 lists some noteworthy audiovisual data resources that were reported in the literature. For each database, we provide the following information: emotion elicitation method (i.e., whether the elicited emotion displays are posed or spontaneous), size (the number of subjects and available data samples), modality (audio and/or visual), emotion description (category or dimension), and labeling scheme. For other surveys of existing databases of human emotion behavior, the readers are referred to Cowie et al. (2005), Pantic et al. (2005), and Zeng, Hu, Liu, Fu, and Huang, (2006).

As far as the databases of deliberate emotion behavior are concerned, the following databases need to be mentioned. The Chen–Huang audiovisual database (Chen, 2000) is to our knowledge the largest multimedia database containing facial and vocal deliberate displays of basic emotions and four cognitive states. The FABO database of Gunes and Piccardi (2006) contains videos of facial expressions and body gestures portraying posed displays of basic and nonbasic emotional states (six prototypical emotions, uncertainty, anxiety, boredom, and neutral).

The existing datasets of spontaneous emotion behavior were collected in one of the following scenarios: human–human conversation (Bartlett et al., 2005; Douglas-Cowie et al., 2003; Roisman et al., 2004), human–computer interaction (SAL), and clips from television (Douglas-Cowie et al., 2003). 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 emotion behavior. In databases of spontaneous emotion behavior, dimensional descriptions in the evaluation-activation space (SAL[2]; Douglas-Cowie et al., 2003), and some

[2] http://emotion-research.net/toolbox/toolboxdatabase.2006-09-26.5667892524.

Table 14.1 Multimedia Databases of Human Emotion Behavior

References	Elicitation Method	Size	Emotion Description	Labeling
FABO face and body gesture (Gunes and Piccardi, 2006)	Posed: two cameras to record facial expressions and body gestures, respectively	23 adults Mixed races Available: 210 videos	Category: 6 basic emotions, neutral, uncertainty, anxiety, boredom	N/A
Chen-Huang '00 (Chen, 2000)	Posed	100 adults, 9900 visual and AV expressions	Category: 6 basic emotions, and 4 cognitive states (interest, puzzle, bore, frustration)	N/A
Adult Attachment Interview '04 (Roisman et al., 2004)	Natural: subjects were interviewed to describe the childhood experience	60 adults Each interview last 30–60min	Category: 6 basic emotions, embarrassment, contempt, shame, general positive and negative.	FACS
RU-FACS '05 (Bartlett et al., 2005)	Natural: subjects were tried to convince the interviewers they were telling the truth	100 adults	Category: 33 AUs	FACS
SAL '05[2]	Induced: subjects interacted with artificial listener with different personalities	24 adults 10 h	Dimensional labeling/categorical labeling	FEEL-TRACE
Belfast database '03 (Douglas-Cowie et al., 2003)	Natural: clips taken from television and realistic interviews with research team	125 subjects. 209 sequences from TV, 30 from interview	Dimensional labeling/categorical labeling	FEEL-TRACE

application-dependent emotional states are usually used as the data labels. Interest, boredom, confusion, frustration, uncertainty, anxiety, embarrassment, contempt, and shame are some examples of the used application-dependent emotion-interpretative labels.

Facial Action Units (AUs) (Ekman et al., 2002) are very suitable to describe the richness of spontaneous facial behavior, as the thousands of anatomically possible facial expressions can be represented as the combination of a few dozens of AUs. Hence, the labeling schemes used to code data include FACS AUs (Roisman et al., 2004; Bartlett et al., 2005) and the Feeltrace system for evaluation-activation dimensional description ((Douglas-Cowie et al., 2003; SAL[2]).

Worthy of mention are the FABO face and body gesture database, SAL database, and Belfast database which are publicly accessible, representing efforts toward enhancing communication and establishing reliable evaluation procedures in this field.

14.3.2 Audiovisual Computing

Influenced by basic emotion theory, most of the existing audiovisual emotion recognition studies investigated recognition of the basic emotions from deliberate displays. Relatively few efforts have been reported toward detection of nonbasic emotional states from deliberate displays. Those include the work of Zeng and his colleagues (Zeng et al., 2004, 2006, 2007b, 2008b), and that of Sebe et al. (2006), who added four cognitive states (interest, puzzlement, frustration, and boredom) considering the importance of these cognitive states in human–computer interaction. A related study conducted on naturalistic data is that of Pal, Iyer, and Yantorno (2006), who designed a system to detect hunger and pain as well as sadness, anger, and fear from infant facial expressions and cries.

Most of the existing methods for audiovisual emotion analysis are based on deliberately posed emotion displays (Go, Kwak, Lee, & Chun, 2003; Busso et al., 2004; Song et al., 2004; Zeng et al., 2004, 2006, 2007b; Wang & Guan, 2005; Hoch, Althoff, McGlaun, & Rigoll, 2005; Sebe et al., 2006). Recently a few exceptional studies have been reported toward audiovisual emotion analysis in spontaneous emotion displays. For example, Fragopanagos et al. (2005), Pal et al. (2006), Caridakis et al. (2006), Karpouzis et al. (2007), and Zeng et al. (2007a) used the data collected in psychological research interviews (Adult Attachment Interview), Pal et al. (2006) used recordings of infant affective displays, whereas Fragopanagos and Taylor (2005), Caridakis et al. (2006), and Karpouzis et al. (2007), used the data collected in *Wizard of Oz* scenarios.

Because the available data were usually insufficient to build a robust machine-learning system for recognition of fine-grained emotional states (e.g., basic emotions), recognition of coarse emotional states was attempted in most of the aforementioned studies. The study of Zeng et al. (2007) focuses on audiovisual recognition of positive and negative emotion, whereas other studies report on classification of audiovisual input data into the quadrants in evaluation-activation space (Fragopanagos et al., 2005; Caridakis et al., 2006; Karpouzis et al., 2007).

The studies reported in Fragopanagos et al., (2005), Caridakis et al. (2006), and Karpouzis et al. (2007) applied the FeelTrace system that enables raters to continuously label changes in emotion expressions. However, note that the study discussed in Fragopanagos et al. (2005) reported on a considerable labeling variation among four human raters due to the subjectivity of audiovisual emotion judgment. More specifically, one of the raters mainly relied on audio information when making judgments whereas 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 study of Zeng et al. (2007) made the

assumption that facial expression and vocal expression have the same coarse emotional states (positive and negative), and they then directly used FACS-based labels of facial expressions as audiovisual expression labels.

The data fusion strategies utilized in the current studies on audiovisual emotion recognition are feature-level, decision-level, or model-level fusion. An example of feature-level fusion is the study in Busso et al. (2004) that concatenated the prosodic features and facial features to construct joint feature vectors which are then used to build an emotion 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 emotion recognizer based on a feature-level fusion.

The vast majority of studies on bimodal emotion recognition reported on decision-level data fusion (Go et al., 2003; Zeng et al., 2004, 2007b; Busso et al., 2004; Hoch et al., 2005; Wang & Guan, 2005; Pal et al., 2006).

In decision-level data fusion, the input coming from each modality is modeled independently and these single-modal recognition results are combined at the end. Because 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, some model-level fusion methods have been proposed that aim at making use of the correlation between audio and visual data streams (e.g., Song et al., 2004; Fragopanagos et al., 2005; Caridakis et al., 2006; Sebe et al., 2006; Zeng et al., 2006, 2008b; Karpouzis et al., 2007). Zeng et al., (2008b) presented multistream 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. (2006) 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. (2004) 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 (2005) proposed an artificial neural network with a feedback loop called ANNA to integrate the information from face, prosody and lexical content. Caridakis et al. (2006) and Karpouzis et al. (2007) investigated combining the visual and audio data streams by using relevant neural networks. Sebe et al. (2006) used a Bayesian network to fuse the facial expression and prosody expression.

14.3.3 Other Multimodal Computing

In addition to the above-mentioned audiovisual emotion recognition studies mainly based on facial expression and prosody expression, a few studies were constructed to investigate the multi-visual-cue fusion, including fusion of facial expressions and

head movements (Cohn et al., 2004; Zhang & Ji, 2005; Ji, Lan, & Looney, 2006), fusion of facial expression and body gesture (Balomenos, Raouzaiou, Ioannou, Drosopoulos, Karpouzis, & Kollias, 2005; Gunes & Piccardi, 2005), fusion of facial expression and gaze (Ji et al., 2006), and fusion of facial expression and head and shoulder movement (Valstar et al., 2007), based on the supplemental contribution of gaze and head and body movement to emotion recognition.

Relatively few reports of automatic emotion recognition are found regarding integration of facial expressions and postures from a sensor chair (Kapoor et al., 2005, 2007), and the integration of facial expression and physiological signals (temperature, heart rate, and skin conductivity) from a sensor mouse (Liao et al., 2006), with the aim of improvement of emotion recognition performance.

With the research shift toward analysis of spontaneous human behavior, analysis of only acoustic information will not suffice for identifying subtle changes in vocal emotion expression. Several audio-based studies investigated the combination of acoustic features and linguistic features (language and discourse) to improve vocal emotion recognition performance. Typical examples of linguistic-paralinguistic-fusion methods are those of Litman & Forbes-Riley (2004) and Schuller, Villar, Rigoll, and Lang (2005), who used spoken words and acoustic features; of Lee and Narayanan, (2005), who used prosodic features, spoken words and information of repetition; of Graciarena, Shriberg, Stolcke, Enos, & Kajarekar (2006), who combined prosodic, lexical, and cepstral features; and of Batliner et al. (2003), who used prosodic features, Part-Of-Speech (POS), Dialogue Act (DA), repetitions, corrections, and syntactic-prosodic boundary to infer the emotion.

Finally, virtually all present approaches to automatic emotion analysis are context-insensitive. Exceptions from this overall state of the art in the field include just a few studies. Pantic & Rothkrantz (2004) investigated interpretation of facial expressions in terms of user-defined interpretation labels. Ji et al. (2006) investigated the influence of context (work condition, sleeping quality, circadian rhythm, and environment, physical condition) on fatigue detection, and Kapoor and Picard (2005) investigated the influence of the task states (difficulty level and game state) on interest detection. Litman et al. (2004) also investigated the role of context information (e.g., subject, gender, and turn-level features representing local and global aspects of the dialogue) on audio emotion recognition.

14.3.4 Exemplar Methods

We introduce in this section our efforts toward machine understanding of multimodal affective behavior, which were published in related conferences and journals.

14.3.4.1 Audiovisual Posed Affective Expression Recognition

The team of Zeng and Huang made efforts (Zeng et al., 2004, 2006, 2007, 2008) toward audiovisual posed affective expression recognition, based on the data of 20 subjects (10 females and 10 males) from the database Chen–Huang (Chen, 2000). Although the subjects displayed affect expressions on request, the subjects chose how to express each state. They were simply asked to display facial expressions and speak appropriate sentences. Each subject was required to repeat each state with speech three times. Therefore, for every affective state, there are $3 * 20 = 60$ video sequences. And there are a total of $60 * 11 = 660$ sequences for 11 affective states. The time of every sequence ranged from 2–6 seconds.

A tracking algorithm called Piecewise Bezier Volume Deformation (PBVD) tracking (Tao & Huang, 1999) was applied to extract facial features in our experiment. This face tracker uses a 3D facial mesh model embedded in multiple Bezier volumes. The tracker can track head motion and local deformations of the facial features. The local deformations of facial features in terms of 12 predefined motions shown in Figure 14.3 are used for affect recognition.

For audio feature extraction, the entropic signal processing system named get_f0, was used to output the pitch F0 for the fundamental frequency estimate, RMS energy for the local root mean squared measurements, prob_voice for the probability of voicing, and the peak normalized cross-correlation value that was used to determine the output F0. The feature selection experimental results in Zeng et al. (2007) showed pitch and energy are the most important audio factors in affect classification. Therefore, in our experiment, we only used these two audio features for affect recognition. Some prosody features, such as frequency and duration of silence, could have implications in the HMM structure of energy and pitch.

For integrating coupled audio and visual features, we present a model-level fusion, called multistream fused HMM (MFHMM) (Zeng et al., 2008b) which devises a new structure linking the multiple component HMMs which is optimal according to the maximum entropy principle and maximum mutual information (MMI) criterion. In the MFHMM framework, the different streams are modeled by different fused HMM components which connect the hidden states of one HMM to the

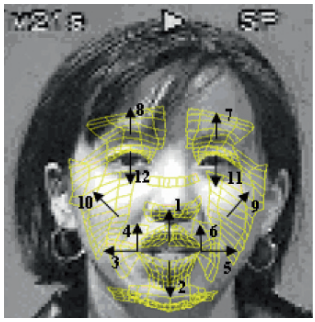

Fig. 14.3 Facial motion units.

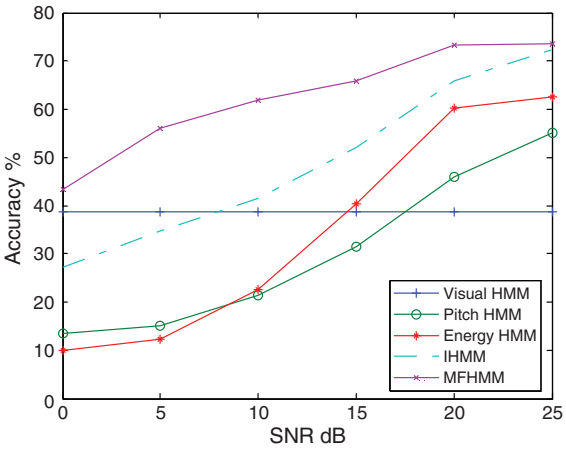

Fig. 14.4 Accuracies of different methods under various audio SNR conditions.

observation of other HMMs. The details of MFHMM, including its learning and inference algorithms, can be found in Zeng et al. (2008b).

We applied leave-one-person-out cross-validation to test our person-independent affect recognition algorithm. In our experiment, the composite facial feature from video, energy, and pitch features from audio are treated as three coupled streams, and modeled by three component HMMs. We used five methods to make decisions: face-only HMM, audio-only HMM, independent HMM fusion (IHMM), and MFHMM. IHMM is multistream HMM that assumes the independence among different stream.

The experimental results under varying levels of audio SNR are shown in Figure 14.4. They demonstrate that audiovisual fusion outperforms unistream methods in most cases; that is, both IHMM and MFHMM are better than visual HMM, pitch HMM, and energy HMM. The exceptions are that the accuracies of IHMM in 0 and 5 dB audio SNR are lower than visual HMM. That shows that the IHMM combination scheme cannot achieve better performance than individual modality when the performance of certain individual streams is very bad. On the other hand, the performance of MFHMM is still a little higher than the visual-only HMM in 0 dB audio SNR. Thus, MFHMM is more robust for processing noisy data than IHMM.

14.3.4.2 Audiovisual Spontaneous Affective Expression Recognition

The team of Zeng and Huang also explored recognition of audiovisual spontaneous affective expressions occurring in a psychological interview named the Adult Attachment Interview (AAI) that is used to characterize individuals' current states of mind with respect to past parent–child experiences. We present the Adaboost multistream hidden Markov model (Adaboost MHMM; Zeng et al., 2007) to integrate audio and visual affective information.

In order to capture the richness of facial expression, we used a 3D face tracker (Tao & Huang, 1999) to extract facial texture images that were then transformed into low-dimensional subspace by locality preserving projection (LPP). We used pitch and energy in the audio channel to build audio HMM in which some prosody features, such as frequency and duration of silence, could have implications. In the audiovisual fusion stage, we treated the component HMM combination as a multiclass classification problem in which the input is the probabilities of HMM components and the output is the target classes, based on the training combination strategy (Zeng et al., 2006). We used the Adaboost learning scheme to build fusion of the component HMMs from audio and visual channels.

Based on the Adaboost learning scheme, these estimated likelihoods of component HMMs were used to construct a strong classifier which is a weighted linear combination of a set of weak classifiers. A set of weaker hypotheses was estimated, each using likelihood of positive or negative emotion of a single-component HMM. The final hypothesis was obtained by weighted linear combination of these hypotheses where the weights were inversely proportional to the corresponding training errors.

The personal-dependent recognition was evaluated on the two subjects (one female and one male). The emotion recognition results of unimodal (audio-only or visual-only) methods and audiovisual fusion are shown in Figure 14.5. Two combination schemes (weighting and training) were used to fuse the component HMMs from the audio and visual channels. Acc MHMM means MHMM with the weighting combination scheme in which the weights are proportional to stream normalized recognition accuracies. Adaboost MHMM means MHMM with the Adaboost learning schemes as described in Section 14.6.

Because we treated the multistream fusion as a multiclass classification problem, there was a variety of methods that could be used to build the fusion. In addition to Adaboost MHMM, we used LDC and KNN approaches to build this audiovisual fusion, which are Ldc MHMM and Knn MHMM in Figure 14.5.

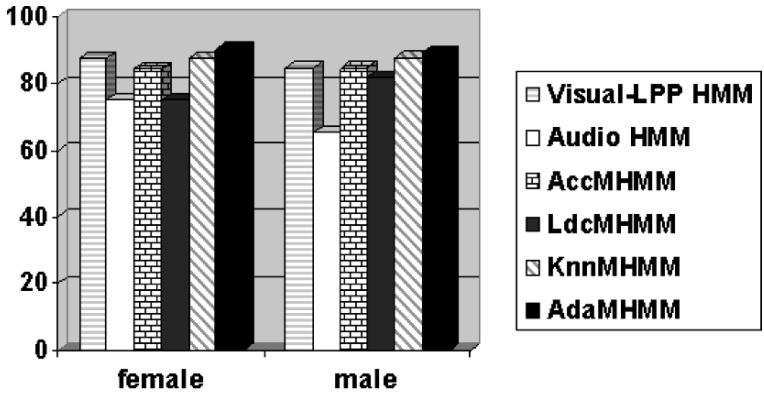

Fig. 14.5 Performance comparison among different modalities and different fusions.

The performance comparison of these fusion methods is as follows: Adaboost MHMM > Knn MHMM > Acc MHMM > Ldc MHMM. The results demonstrate that the training combination outperforms the weighting combination, except Ldc MHMM which is a linear fusion. Adaboost MHMM is the best among these four fusion methods. Results show that Adaboost MHMM and Knn MHMM are better than unimodal HMM (i.e., visual-only HMM and audio-only HMM). That suggests that multiple modalities (audio and visual modalities) can provide more affective information and have the potential to obtain better recognition performance than a single modality.

In Figure 14.5, the accuracy of Acc MHMM equals the visual-only HMM for male data but is worse than visual-only HMM for female data. Ldc MHMM is worse than visual-only HMM in female and male cases. Both Acc MHMM and Ldc MHMM are linear bimodal fusion. That suggests that the fusion method plays an important role in audiovisual emotion recognition. Although the linear bimodal combination is reasonable and intuitive, it is not guaranteed to obtain the optimal combination in a realistic application. It is even possible that this combination is worse than individual component performance, as shown in our experiments.

14.3.4.3 Multi-Visual-Cue Analysis of Posed Versus Spontaneous Expressions

Pantic's team proposed a system of automatic discrimination between posed and spontaneous smiles based on multiple visual cues including facial expression (described in terms of AUs that produce it) and head and shoulder movements (Valstar et al., 2007). Although few systems have been recently reported on automatic discrimination between spontaneous and posed facial behavior (i.e., Valstar et al., 2006; Littlewort et al., 2007), this is the only reported effort so far to automatically discern spontaneous from deliberately displayed behavior based on multiple visual cues.

Conforming with the research findings in psychology, the method relies on characteristics of temporal dynamics of facial, head, and shoulder actions and employs parameters such as speed, intensity, duration, and occurrence order of visual behavioral cues to classify smiles present in an input video as either deliberate or genuine smiles.

A cylindrical head tracker, proposed by Xiao et al. (2003), was used to track head motion; auxiliary particle filtering tracking method, proposed by Pitt and Shephard (1999), was used to track the tip points of the shoulders; and particle filtering with factorized likelihood, proposed by Patras and Pantic (2004), was used to track 20 facial characteristic points such as the corners of the mouth and the eyes (see Figure 14.6). Using the tracking data, the presence (i.e., activation) of AU6 (raised cheeks), AU12 (lip corners pulled up), AU13 (lip corners pulled up sharply), head movement (moved off the frontal view), and shoulder movement (moved off the relaxed state), were detected first. For each of these behavioral cues, the temporal segments (neutral, onset, apex, and offset) were also determined.

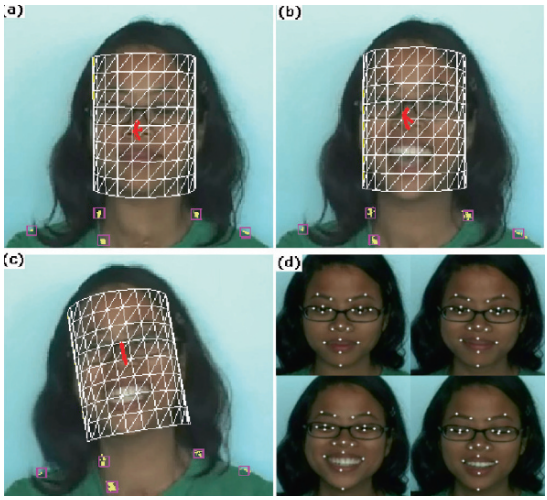

Fig. 14.6 Illustration of the tracking procedure used by Valstar et al. (2007).

To detect the activated AUs and their temporal segments, the AU detector proposed by Valstar and Pantic (2006), which combines Gentle Boost ensemble learning and Support Vector Machines, was used. To detect head and shoulder actions and their temporal segments a rather simple rule-based expert system was used. Then a set of midlevel feature parameters was computed for every temporal segment of each present behavioral cue. These parameters included the segment duration, the mean and the maximum displacements of the relevant points in *x*- and *y*-directions, the maximum velocity in *x*- and *y*-directions, and the asymmetry in the displacements.

In addition, the second-order polynomial functional representation of the displacements of the relevant points, and the order in which the behavioral cues have been displayed, were computed as well. Gentle Boost has been used to learn the most informative parameters for distinguishing between spontaneous and volitional smiles and these parameters have been used further to train a separate Support Vector Machine for each temporal segment of each of the five behavioral cues (i.e., in total 15 GentleSVMs). The outcomes of these 15 GentleSVMSs are then combined and a probabilistic decision function determines the class (spontaneous or posed) for the entire smile episode. When trained and tested on a set containing 100 samples of volitional smiles and 102 spontaneous smiles from the MMI Facial Expression database (Pantic et al., 2005, Pantic & Bartlett, 2007), the proposed method attained a 94% correct recognition rate (0.964 recall with 0.933 precision) when determining the class (spontaneous or posed) of an input facial expression of smile.

14.3.4.4 Audiovisual Laughter Detection

Research in cognitive sciences provided some promising hints that vocal outbursts and nonlinguistic vocalizations such as yelling, laughing, and sobbing may be very important cues for decoding someone's affect/attitude (Russell et al., 2003),

therefore few efforts toward automatic recognition of nonlinguistic vocal outbursts have been recently reported (Zeng et al., 2008a). Most of these efforts are based only on audio signals such as the method for automatic laughter detection proposed by Truong and van Leeuwen, (2007).

However, because it has been shown by several experimental studies in either psychology or signal processing that integrating the information from audio and video leads to an improved performance of human behavior recognition (e.g., Petridis and Pantic (2008)), few pioneering efforts toward audiovisual recognition of nonlinguistic vocal outbursts have been recently reported including audiovisual analysis of infants' cries proposed by Pal et al. (2006) and audiovisual laughter recognition proposed by Petridis and Pantic (2008). The latter is the only effort reported so far aimed at automatic discrimination between laughter and speech in naturalistic data based on both facial and vocal expression.

The method uses particle filtering with factorized likelihood, proposed by Patras and Pantic 2004), to track 20 facial characteristic points such as the corners of the mouth and the eyes in an input video. Then it uses principal component analysis (PCA) to distinguish changes in the location of facial points caused by changes in facial expression from those caused by rigid head movements. The changes in facial expression are used in further processing. To detect laughter from the audio signal, spectral features, namely perceptual linear prediction coding features and their temporal derivatives, have been used.

Both feature-level fusion and decision-level fusion have been investigated. To achieve decision-level fusion, the input coming from each modality (audio and video) was modeled independently by a neural network preceded by Ada Boost, which learns the most informative features for distinguishing between laughter and speech as shown by the relevant modality. Then, these single-modal recognition results were combined at the end using the SUM function. To achieve feature-level fusion, audio and video features were concatenated into a single feature vector (by up-sampling so that the video frame rate equaled the audio frame rate of 50 frames per second).

The resulting feature vector was then used to train the target Ada neural network. When trained and tested on a set of 40 audiovisual laughter segments and 56 audiovisual speech segments from the AMI corpus (http://corpus.amiproject.org/), the proposed method attained an average 86% correct recognition rate (feature-level fusion: 0.869 recall with 0.767 precision; decision-level fusion: 0.817 recall with 0.823 precision) when determining the class (speech or laughter) of an input audiovisual episode.

14.4 Challenges

There are two new trends in the research on automatic human emotion recognition: analysis of spontaneous emotion behavior and multimodal analysis of human emotion behavior including audiovisual analysis, combined audio-based linguistic

and nonlinguistic analysis, and multicue visual analysis based on facial expressions, gaze, 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., face, gaze, head, and body gestures, and linguistic and paralinguistic information). This section discusses these issues.

14.4.1 Databases

Acquiring valuable spontaneous emotion behavior data and the related ground truth is far from being solved. To our knowledge, there is still no database of emotion behavior that records data of all emotion-related modalities (facial and vocal expression, physiological responses, body movement, and gaze). In addition, although much effort has been made toward collection of audiovisual databases of spontaneous human emotion behavior, most of the data contained in the available databases currently lack labels. In other words, no metadata are available that could identify the emotional state displayed in a video sample and the context in which this emotional state was displayed.

Although some promising coding systems have been proposed to label the facial action (e.g., FACS (Ekman et al., (2002)), and audiovisual expression (e.g., Feeltrace (Cowie et al. 2000)), how to reliably code all emotion-related modalities for automatic emotion systems remains an open issue. Much work is needed to understand the correlation of the dynamic structure of these modalities. The metadata about the context in which the recordings were made such as the utilized stimuli, the environment, and the presence of other people, are needed because these contextual variables may influence masking of the emotional reactions. In addition, human labeling of emotion behavior is very time consuming and full of ambiguity due to human subjective perception. Improving the efficiency and reliability of labeling multimodal emotion databases is critical in the recognition of spontaneous emotion.

Readers are referred to Pantic and Rothkrantz (2003), Pantic et al. (2005), Cowie et al. (2005), and Zeng et al., (2006), which discussed a number of specific research and development efforts needed to build a comprehensive, readily accessible reference set of emotion displays that could provide a basis for benchmarks for all different efforts in the research on machine analysis of human emotion behavior.

14.4.2 Multimodal Computing

Although 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 emotion, it remains unclear how this should be accomplished. Studies in neurology on fusion of sensory neurons (Stein & Meredith, (1993) 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 emotion 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 the fields of machine learning and pattern recognition. Many studies have demonstrated the advantage of classifier fusion over the individual classifiers due to the uncorrelated errors from different classifiers (e.g., Kuncheva, 2004). Various classifier fusion methods (fixed rules and trained combiners) have been proposed in the literature, but optimal design methods for classifier fusion are still not available. In addition, because humans simultaneously employ tightly coupled audio and visual modalities, 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 combining the benefits of both feature-level and decision-level fusion methods may be a good choice for this fusion problem. However, based on existing knowledge and methods, how to model multimodal fusion for spontaneous emotion displays 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 the 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 the interpretation of human naturalistic, multimodal emotion behavior. Yet these are virtually unexplored areas of research, due to the fact that facial expression and vocal expression of emotion are usually studied separately.

An important related issue that should be addressed in multimodal emotion recognition is how to make use of information about the context (environment, observed subject, his or her current task) in which the observed emotion behavior was displayed. Emotions are intimately related to a situation being experienced or imagined by a human. Without context, a human may misunderstand the observed person's emotion expressions. Yet, with the exception of a few studies investigating the influence of context on emotion recognition (e.g., Litman & Forbes-Riley, 2004; Pantic & Rothkrantz, 2004; Kapoor & Picard, 2005; Kapoor et al., 2007; Ji et al., 2006), virtually all existing approaches to machine analysis of human emotion are context-insensitive.

Building a context model that includes person ID, gender, age, conversation topic, and workload needs help from other research fields such as face recognition, gender recognition, age recognition, topic detection, and task tracking. Because the problem of context sensing is very difficult to solve, pragmatic approaches (e.g., activity- and/ or subject-profiled approaches) should be taken.

14.5 Conclusion

Multimodal information fusion is a process that enables human ability to assess emotional states robustly and flexibly. In order to understand the richness and subtlety of human emotion behavior, the computer should be able to integrate all emotion-related modalities (facial expression, speech, body movement, gaze, and physiological response). Multimodal emotion recognition has received rather less research attention than single modality, due to the complexity of multimodal processing and the availability of relevant training and test material. As this chapter describes, we have witnessed increasing efforts toward machine analysis of multimodal emotion recognition in the last decade. Specifically, some databases of acted and spontaneous emotion displays have been collected by researchers in the field, and a number of promising methods for multimodal analysis of human spontaneous behavior have been proposed.

In the meantime, several new challenging issues have been identified. Let us reconsider the interview scenario described in the introduction where the subject expressed her emotion by using facial expression, speech (linguistic content and paralinguistic behavior), head movement, and gaze in an efficient way. The complexity of the occurring emotion behavior challenges our current knowledge and approaches in machine analysis of emotion behavior.

In order to enable the computer to reliably perceive human emotion displays as humans do, much effort across multiple research disciplines is needed to address the following important issues: build a comprehensive, readily accessible reference set of emotion displays that could provide a basis for benchmarks for all different efforts in the field; develop methods for spontaneous emotion behavior analysis that are robust to the observed person's arbitrary movement, occlusion, complex and noisy background; devise models and methods for human emotion analysis that take into consideration temporal structures of the modalities and temporal correlations between the modalities (and/or multiple cues), and context (subject, his or her task, environment); and develop better methods for multimodal fusion.

The progress of research on automatic emotion recognition based on multimodal behavior and physiological components will greatly contribute to developing a comprehensive understanding of the mechanisms involved in human behavior, and enhancing human–computer interaction through emotion-sensitive and socially intelligent interfaces.

Acknowledgment We would like to thank Professor Glenn I. Roisman for providing the valuable videos and FACS codes of the Adult Attachment Interview. This work of Zhihong Zeng and Thomas Huang was supported in part by the Beckman Postdoctoral Fellowship and NSF CCF 04-26627. The work of Maja Pantic has been funded in part by EU IST Program Project FP6-0027787 (AMIDA) and EC's 7th Framework Program [FP7/2007-2013] under grant agreement no. 211486 (SEMAINE).

References

Adams, R. B & Kleck, R.E(2003). Perceived gaze direction and the processing of facial displays of emotion. *Psychological Science, 14*, 644–647.

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*(2), 256–274.

Balomenos, T., Raouzaiou, A., Ioannou, S., Drosopoulos, A., Karpouzis, K., & Kollias, S. (2005). *Emotion analysis in man-machine interaction systems* (LNCS 3361; pp. 318–328). New York: Springer.

Bartlett, M. S., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., & Movellan, J. (2005), Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 568–573).

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

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. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 205–211).

Caridakis, G., Malatesta, L., Kessous, L., Amir, N., Paouzaiou, A. & Karpouzis, K. (2006). Modeling naturalistic affective states via facial and vocal expression recognition. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 146–154).

Chen, L., Huang, T. S., Miyasato, T., & Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition* (pp. 396–401).

Chen, L. S. (2000), Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction, PhD thesis, University of Illinois at Urbana-Champaign, USA.

Cohn, J. F. (2006). Foundations of human computing: Facial expression and emotion. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 233–238).

Cohn, J. F., Reed, L. I., Ambadar, Z., Xiao, J., & Moriyama, T. (2004). Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior. In *Proceedings of the International Conference on Systems, Man & Cybernetics, 1* (pp. 610–616).

Cohn, J. F., & 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.

Cowie, R., Douglas-Cowie, E., & Cox, C. (2005). Beyond emotion archetypes: Databases for emotion modeling using neural networks. *Neural Networks, 18*, 371–388.

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

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. G. (2001), Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine,* January (pp. 32–80).

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of database. *Speech Communication, 40*(1–2), 33–60.

Duric, Z., Gray, WḊ., 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, 90*(7), 1272–1289.

Ekman, P. (Ed.) (1982). *Emotion in the human face* (2nd ed.). New York: Cambridge University Press.

Ekman, P., & Friesen, W. V. (1975). *Unmasking the face*. Englewood Cliffs, NJ: Prentice-Hall.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). *Facial Action Coding System*. Salt Lake City, UT: A Human Face.

Ekman P., & Rosenberg, E. L. (2005). What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (2nd ed.). Oxford University Press, University of Illinois at Urbana-Champaign, USA.

Fragopanagos, F., & Taylor, J. G. (2005), Emotion recognition in human–computer interaction. *Neural Networks, 18*, 389–405.

Go, H. J., Kwak, K. C., Lee, D. J., & Chun, M.G. (2003). Emotion recognition from facial image and speech signal. In *Proceedings of the International Conference of the Society of Instrument and Control Engineers* (pp. 2890–2895).

Graciarena, M., Shriberg, E., Stolcke, A., Enos, J. H. F., & Kajarekar, S. (2006). Combining prosodic, lexical and cepstral systems for deceptive speech detection. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, I*, 1033–1036.

Gunes, H., & Piccardi, M. (2005). Affect recognition from face and body: early fusion vs. late fusion. In *Proceedings of the International Conference on Systems, Man and Cybernetics* (pp. 3437–3443).

Gunes, H., & Piccardi, M. (2006). A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. *International Conference on Pattern Recognition, 1*, 1148–1153.

Harrigan, J. A., Rosenthal, R., & Scherer, K. R. (2005). *The new handbook of methods in nonverbal behavior research* (pp. 369–397). Oxford University Press, USA.

Hoch, S., Althoff, F., McGlaun, G., & Rigoll, G. (2005), Bimodal fusion of emotional data in an automotive environment. In *ICASSP, II* (pp. 1085–1088).

Ji, Q., Lan, P., & Looney, C. (2006). A probabilistic framework for modeling and real-time monitoring human fatigue. *IEEE SMC-Part A, 36*(5), 862–875.

Kapoor, A., Burleson, W., & Picard, R. W. (2007). Automatic prediction of frustration. *International Journal of Human–Computer Studies, 65*(8), 724–736.

Kapoor, A., & Picard, R. W. (2005). Multimodal affect recognition in learning environment. In *ACM International Conference on Multimedia* (pp. 677–682).

Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L., & Kollias, S. (2007). Modeling naturalistic affective states via facial, vocal, and bodily expression recognition (LNAI 4451; pp. 91–112). New York: Springer.

Kuncheva, L. I. (2004). *Combining pattern classifier: Methods and algorithms*. Hoboken, NJ: John Wiley and Sons.

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

Liao, W., Zhang, W., Zhu, Z., Ji, Q., & Gray, W. (2006), Toward a decision-theoretic framework for affect recognition and user assistance. *International Journal of Human-Computer Studies, 64*(9), 847–873.

Lisetti, C. L., & Nasoz, F. (2002). MAUI: A multimodal affective user interface. In *Proceedings of the International Conference on Multimedia* (pp. 161–170).

Lisetti, C. L., & Nasoz, F. (2004). Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing, 11*, 1672–1687.

Litman, D. J., & Forbes-Riley, K. (2004), Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, July (pp. 352–359).

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

Maat, L., & Pantic, M. (2006). Gaze-X: Adaptive affective multimodal interface for single-user office scenarios, In *Proceedings of the ACM International Conference on Multimodal Interfaces* (pp. 171–178).

Pal, P., Iyer, A. N., & Yantorno, R. E. (2006). Emotion detection from infant facial expressions and cries. In *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, 2* (pp. 721–724).

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

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. S. (2006). Human computing and machine understanding of human behavior: A survey. In *International Conference on Multimodal Interfaces* (pp. 239–248).

Pantic M., & Rothkrantz, L. J. M. (2003), Toward an affect-sensitive multimodal human-computer interaction, *Proceedings of the IEEE, 91*(9, Sept.), 1370–1390.

Pantic, M., & Rothkrantz, L. J. M. (2004). Case-based reasoning for user-profiled recognition of emotions from face images. In *International Conference on Multimedia and Expo* (pp. 391–394).

Pantic, M., Valstar, M. F, Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In *International Conference on Multimedia and Expo* (pp. 317–321).

Patras, I., & Pantic, M. (2004). Particle filtering with factorized likelihoods for tracking facial features, In *Proceedings of the IEEE International Conference on Face and Gesture Recognition* (pp. 97–102).

Pentland, A. (2005). Socially aware, computation and communication, *IEEE Computer, 38*, 33–40.

Petridis, S., & Pantic, M. (2008). Audiovisual discrimination between laughter and speech, In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing* (pp. 5117–5120).

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(10), 1175–1191.

Pitt, M. K., & Shephard, N. (1999). Filtering via simulation: auxiliary particle filtering. *Journal of the American Statistical Association, 94*, 590–599.

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, 40*(5), 776–789.

Russell, J. A., Bachorowski, J., & Fernandez-Dols, J. (2003). Facial and vocal expressions of emotion. *Ann. Rev. Psychol. 54*, 329–349.

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

Schuller, B., Villar, R. J., Rigoll, G., & Lang, M. (2005). Meta-classifiers in acoustic and linguistic feature fusion-based affect recognition. In *International Conference on Acoustics, Speech, and Signal Processing* (pp. 325–328).

Sebe, N., Cohen, I., Gevers, T., & Huang, T. S. (2006). Emotion recognition based on joint visual and audio cues. In *International Conference on Pattern Recognition* (pp. 1136–1139).

Sebe, N., Cohen, I., & Huang, T. S. (2005). Multimodal emotion recognition. In *Handbook of Pattern Recognition and Computer Vision*. Singapore: World Scientific.

Song, M., Bu, J., Chen, C., & Li, N. (2004), Audio-visual based emotion recognition—A new approach. In *International Conference on Computer Vision and Pattern Recognition* (pp. 1020–1025).

Stein, B., & Meredith, M. A. (1993). *The merging of senses*. Cambridge, MA: MIT Press.

Stemmler, G. (2003). Methodological considerations in the psychophysiological study of emotion. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 225–255). Oxford University Press, USA.

Tao, H., & Huang, T. S. (1999), Explanation-based facial motion tracking using a piecewise Bezier volume deformation mode. In *CVPR'99, 1* (pp. 611–617).

Truong, K. P., & van Leeuwen, D. A. (2007) Automatic discrimination between laughter and speech, *Speech Communication, 49*, 144–158.

Valstar, M. F., Gunes, H., & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. *In ACM Int'l Conf. Multimodal Interfaces* (pp. 38–45).

Valstar, M., Pantic, M., Ambadar, Z., & Cohn, J. F. (2006). Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *International Conference on Multimedia Interfaces* (pp. 162–170).

Valstar, M. F., & Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, 3*,149.

Wang, Y., & Guan, L. (2005), Recognizing human emotion from audiovisual information. In *ICASSP, II* (pp. 1125–1128).

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

Xiao, J., Moriyama, T., Kanade, T., & Cohn, J. F. (2003). Robust full-motion recovery of head by dynamic templates and re-registration techniques. *International Journal of Imaging Systems and Technology, 13*(1), 85–94.

Yoshimoto, D., Shapiro, A., O'Brian, K., & Gottman, J. M. (2005). Nonverbal communication coding systems of committed couples. In *New handbook of methods in nonverbal behavior research*, J.A. Harrigan, R. Rosenthal, and K. R. Scherer (Eds.) (pp. 369–397), USA.

Zeng, Z., Hu, Y., Liu, M., Fu, Y., & Huang, T. S.(2006), Training combination strategy of multi-stream fused hidden markov model for audio-visual affect recognition. In *Proceedings of the ACM International Conference on Multimedia* (pp. 65–68).

Zeng, Z., Hu, Y., Roisman, G. I., Wen, Z., Fu, Y., & Huang, T. S. (2007a), Audio-visual spontaneous emotion recognition. In T. S. Huang, A. Nijholt, M. Pantic, & A. Pentland (Eds.) *Artificial Intelligence for Human Computing* (LNAI 4451, pp. 72–90). New York, Springer.

Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2008a). A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (in press).

Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T. S., Roth, D., & Levinson, S. (2004), Bimodal HCI-related Emotion Recognition, In *International Conference on Multimodal Interfaces* (pp. 137–143).

Zeng, Z., Tu, J., Pianfetti, B., & Huang, T. S. (2008b). Audio-visual affective expression recognition through multi-stream fused HMM. *IEEE Transactions on Multimedia*, June 2008, *10*(4), 570–577.

Zeng, Z., Tu, J., Liu, M., Huang, T. S., Pianfetti, B., Roth, D., & Levinson, S. (2007b). Audio-visual affect recognition. *IEEE Transactions on Multimedia, 9*(2), 424–428.

Zhang, Y., & Ji, Q. (2005). Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(5), 699–714.