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

Zhihong Zeng[1], Maja Pantic[2], Glenn I. Roisman[1] and Thomas S. Huang[1]
[1]University of Illinois at Urbana-Champaign, USA
[2]Imperial College London, UK / University of Twente, Netherlands
{zhzeng,huang}@ifp.uiuc.edu, m.pantic@imperial.ac.uk, roisman@uiuc.edu

ABSTRACT
Automated analysis of human affective behavior has attracted increasing attention from researchers in psychology, computer science, linguistics, neuroscience, and related disciplines. Promising approaches have been reported, including automatic methods for facial and vocal affect recognition. However, the existing methods typically handle only deliberately displayed and exaggerated expressions of prototypical emotions--despite the fact that deliberate behavior differs in visual and audio expressions from spontaneously occurring behavior. Recently efforts to develop algorithms that can process naturally occurring human affective behavior have emerged. This paper surveys these efforts. We first discuss human emotion perception from a psychological perspective. Next, we examine the available approaches to solving the problem of machine understanding of human affective behavior occurring in real-world settings. We finally outline some scientific and engineering challenges for advancing human affect sensing technology.

Categories and Subject Descriptors
A.1 [**Introduction and Survey**]
H.1.2 [**User/Machine Systems**]: Human information processing
H.5.1 [**Multimedia Information Systems**]: Evaluation/ methodology
I.5.4 [**Pattern Recognition Applications**]

General Terms
Algorithms, Performance.

Keywords
Multimodal human computer interaction, multimodal user interfaces, affective computing, human computing, affect recognition, emotion recognition.

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 [15], [20], [57], [63], [64]. 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 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. Human Computing paradigm suggests that user interfaces of the future need to be proactive and human-centered, based on naturally occurring multimodal human communication [57]. More 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.

Fig 1 illustrates a prototype of such an affect-sensitive, multimodal computer-aided learning system. The system was built during the NSF ITR project titled "Multimodal Human Computer Interaction: Toward a Proactive Computer"[1]. In this learning environment, the user explores Lego gear games by interacting with a computer avatar. Multiple sensors are used to detect and track the user's behavioral cues and his or her task. More specifically, the useful information recognized from these sensors includes the user's emotional state, engagement state, the utilized speech keywords, and the gear state. Based on this information, the avatar offers an appropriate tutoring strategy in this interactive learning environment. Other examples of affect-sensitive, multimodal HCI systems include the system of Duric et al. [22], 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, and the proactive HCI tool of Maat and Pantic [51] capable of learning the user's context-dependent behavioral patterns from multi-sensory data and of adapting the interaction accordingly, and the automated Learning Companion of Kapoor et al. [43] that combines information from cameras, a sensing chair and mouse, and wireless skin sensor to detect frustration in order to predict when the user need help. These systems demonstrate a rough picture of future multimodal human-computer interaction.

Except in standard HCI scenarios, potential commercial applications of automatic human affect recognition include affect-sensitive systems for customer services, call centers [46], intelligent automobile system [40], and game and entertainment industry. These systems will change the nature of human-computer interaction in our daily lives. Another important

ICMI'07, November 12–15, 2007, Nagoya, Aichi, Japan.

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

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 [27], [66].

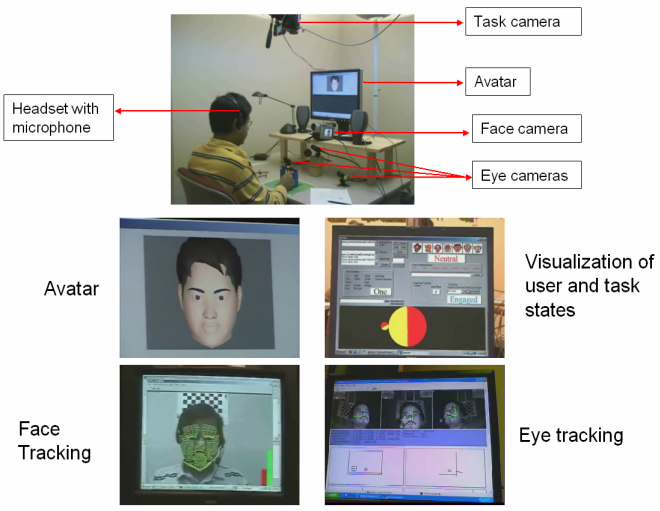

Fig. 1. A prototype of multimodal computer-aided learning system

Because of this practical importance and the theoretical interest of cognitive scientists, automatic human affect analysis has attracted the interest of many researchers. However, most of the existing approaches to automatic human affect analysis are uni-modal (e.g., visual-only or audio-only) approaches, based on deliberately displayed affective expressions, and aimed at prototypical (basic) emotions. Accordingly, the efforts toward uni-modal analysis of artificial affective expressions have been focused in the previously published survey papers [20], [55], [57], [58], [59], [61], [69], [75] among which the papers of Cowie et al. in 2001 [20] and of Pantic and Rothkrantz in 2003 [59] have been most comprehensive and widely cited in this field to date.

Due to the criticisms received from both cognitive and computer scientist that the existing methods for automatic human affect analysis are not applicable in real-life situations, where subtle changes in expressions typify the displayed affective behavior rather than the exaggerated changes that typify posed expressions, the focus of the research in the field has started to shift to automatic analysis of spontaneously displayed affective behavior, i.e., spontaneous facial expressions (e.g., [5], [15], [70], [78]) and audio expressions (e.g., [7], [46]). In addition, more and more researchers realize that integrating the information from audio and visual channels leads to an improved recognition of affective behavior occurring in real-world settings. As a result, an increased number of studies on audiovisual human affect recognition have emerged in recent years (e.g., [10], [30], [86]).

This paper introduces and surveys these recent advances in the research on human affect recognition. In contrast to those previous survey papers in the field, it focuses on the approaches that can handle audio and/or visual recordings of spontaneous (as opposed to posed) displays of affective states.

It is organized as follows. Section 2 describes human perception of affect from a psychological perspective. Section 3 provides a detailed review of related studies, specifically available audio/visual computing methods. Section 4 discusses the challenges in enhancing and extending these reviewed studies. A summary and closing remarks conclude the paper.

2. HUMAN AFFECT (EMOTION) PERCEPTION

Constructing an affect analyzer is dependent on our understanding of the nature of affect. This knowledge of affect includes the description of affect, and the association between observed signals (audio and visual signals in this paper) and affective states. There is no doubt that the progress in automatic affect recognition is in part contingent on the progress of psychologists' and linguists' understanding of human affect perception [26], [67].

2.1 The Description of Affect

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 [20], [26], [67]. The most popular example of this description is the prototypical (basic) emotion categories, which include happiness, sadness, fear, anger, disgust, and surprise. The description of basic emotions was supported especially by the cross-cultural studies conducted by Ekman [23]. This influence of basic emotion theory resulted in the fact that most of existing studies of automatic affect recognition focus on recognizing these basic emotions. However, discrete lists of emotions fail to describe the range of emotions occurring in natural communication settings. In particular, basic emotions cover a rather small part of our daily emotional displays. Selection of affect categories that people show in daily interpersonal interactions needs to be done in a pragmatic and context-dependent manner.

An alternative to category description is the dimensional description [20], [32] where an affective state is represented as a point of a set of dimensions defined by psychological concepts. One of the popular methods to describe affective is in terms of dimensions of evaluation and activation [20]. 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 category representation, dimensional representation enables raters to label a range of emotions. However, this 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). Some studies [33] use the additional dimension (e.g., dominance) to add discriminability of emotions.

2.2 Association Between Affects, Audio and Visual Signals

The face plays a significant role in human emotion perception and expression. The association between face and affective arousal was confirmed by a series of impressive and systematic studies in the field of psychology [26], [67].

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 [15]. The aim of sign judgment is to describe the appearance rather than meaning of the shown behavior. While message judgment is focused on interpretation, sign judgment attempts to be objective, leaving the inference about the conveyed message to higher order decision making. The 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. [25]. 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). These AUs can be used for any higher order decision making process including recognition of basic emotions according to Emotional FACS (EMFACS) rules[2] and a variety of affective states according to FACS Affect Interpretation Database (FACSAID)[2], as well as for recognition of other complex psychological states such as depression [27] or pain [49]. 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 higher-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 aimed at automatic AU recognition (e.g., [5], [16], [78]).

Speech is another important communication device in human communication. It delivers affective information through explicit (linguistic) message, and implicit (paralinguistic) message that reflects the way the words are spoken. Although cognitive scientists have not identified the optimal set of vocal cues that reliably discriminate among affective and attitudinal states, listeners seem to be rather accurate in decoding some basic emotions from prosody [41] and some non-basic affective states such as distress, anxiety, boredom, and sexual interest from nonlinguistic vocalizations like laughs, cries, sighs, and yawns [67]. The basic-emotion-related prosodic features extracted from audio signal include pitch, energy, and speech rate. Cowie et al. [20] provided a comprehensive summary of qualitative acoustic correlations for prototypical emotions.

Linguistic content of speech definitely carries emotional information. Some of this information can be inferred directly from the surface features of words which were summarized in some affective word dictionaries and lexical affinity [80], [65]. The rest of this information lies below the text surface and can only be detected when the semantic context (e.g., discourse information) is taken into account. The association between linguistic content and emotion is language-dependent and generalizing from one language to another is very difficult to achieve.

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., [26], [67]). Ekman [24] found that the relative contributions of facial expression, speech and body cues to affect judgment depend both on the affective state and the environment where the affective behavior occurs. Many studies indicate that the human judgment agreement is typically higher for facial expression modality than it is for vocal expression modality. 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., [15], [27], [67]). 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., [15], [27], [78]) as well as for categorization of complex behaviors like pain, shame, and amusement (e.g., [27]). 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 affective behavior. However, these are largely unexplored areas of research. Another 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 [67].

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 address the problem of automatic analysis of spontaneous affective behavior recorded in real-world settings. Keeping in mind the complexity of affective computing, we also briefly examine studies that represent exemplary approaches to treating a specific problem relevant for advancing human affect sensing technology.

For exhaustive surveys of the past efforts in the field, readers are referred to [20], [55], [57], [58], [59], [61], [69], [75].

This section is focused on an overview of the existing computing methods for automatic human affect recognition based on audio and/or visual displays. For the surveys of existing databases of spontaneous human affective behavior, the readers are referred to [18], [34], [62].

3.1 Facial Expression Recognition

The current research of facial expression recognition can be divided into two directions [15]: recognition of affect and recognition of facial muscle action (facial action units).

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 materials (e.g., [42]). There are a few tentative efforts to detect non-basic affective states from deliberately displayed facial expressions including fatigue [40], pain [49], and mental states like agreeing, concentrating, disagreeing, interest, frustration, thinking and unsure [28], [43], [82].

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

Growing efforts are recently reported toward automatic analysis of spontaneous facial expression data [5], [6], [15], [16], [17], [39], [49], [50], [70], [78], [84]. Some of them study automatic recognition of AUs rather than emotions from spontaneous facial displays [5], [6], [15], [16], [78]. Several of these studies [17], [78] investigated the difference between spontaneous and deliberate facial behavior. The study [17] showed that 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. In addition, it has been shown in [78] 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.

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. [13], who used a shape model defined by 58 facial landmarks, and of Pantic and her colleagues [56], [60], [78], who used a set of facial characteristic points around the mouth, eyes, eyebrows, nose, and chin. Typical example of hybrid, geometric- and appearance-feature-based method, is that of Zhang and Ji [90], who used 26 facial points around the eyes, eyebrows, and mouth and the transient features like crow-feet wrinkles and nasal-labial furrows. Typical examples of appearance-feature-based methods are those of Bartlett et al. [5], [6] and Guo and Dyer [36], who used Gabor wavelets or eigenfaces, of Anderson and McOwen [1], who used a holistic spatial ratio face template, of Valstar et al. [77], who used temporal templates, and of Chang et al. [11], who built a probabilistic recognition algorithm based on the manifold subspace of aligned face appearances. An exemplar method of using both geometric and appearance features is that proposed by Lucey et al. [50], that 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 exception is the study of Pantic and Patras [56], who have explored automatic analysis of facial expressions from the profile-view of the face.

Few approaches to automatic facial expression analysis are based on 3D face models. Huang and his colleagues (i.e., [14], [70], [84]) used the geometry or appearance features extracted by a 3D face tracker called Piecewise Bezier Volume Deformation Tracker [74]. Cohn et al. [16] focused on analysis of brow action units and head movement based on a cylindrical head model [81]. Chang et al. [12] and Yin et al. [83] 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.

Relatively few studies investigated the fusion of the information from facial expressions and head movements [16], [40], [90], and the fusion of facial expression and body gesture [4], [35], [43], with the aim to improve affect recognition performance. Except for few studies, e.g., the studies [60], [29] that investigated interpretation of facial expressions in terms of user-defined interpretation labels, and the study [40] that investigated the influence of context (work condition, sleeping quality, circadian rhythm, and environment, physical condition) on fatigue detection, the existing automatic facial expression analyzers are context insensitive.

3.2 Audio Expression Recognition

Research on audio expression recognition is also influenced by basic emotion theory so that most of the existing efforts toward this direction chose the basic emotions or a subset of them as recognized targets. There are a few tentative studies that have investigated the detection of certain application-dependent affective states. Examples of these studies are those of Hirschberg et al. [37], who attempted deception detection, of Liscombe et al. [47], who focused on detecting certainness, Kwon et al. [45], who focused on detecting stress, of Zhang et al. [89], who focused on detecting confidence, confusion, and frustration, of Batliner et al. [7], who focused on detecting trouble, of Ang et al. [2], who focused on detecting annoyance and frustration, and of Steidl et al. [71], who conducted detection of motherese and empathy. More recently, few efforts towards automatic recognition of nonlinguistic vocalizations like laughters [76] and cries [54] have also been reported.

Some researchers started to turn their focus to investigation of spontaneous emotion recognition by using the audio data collected in call centers [46], [52], meetings [52], wizard of OZ [7] or other dialogue systems [8], [48]. In this natural interaction data, affective expressions are often subtle, and basic emotion expressions seldom occurred. Accordingly, these studies always chose to detect coarse affective states, i.e., positive, negative and neutral in [46], [52], [48], or application-dependent states as described above.

When the research shifts from posed emotion expression to spontaneous emotion expression, only acoustic information is not enough to detect the change of audio affective expression, as indicated by Batliner et al. [7] that "the closer we get to a realistic scenario, the less reliable is prosody as an indicator of the speakers emotional state". Thus, a few studies investigated the combination of acoustic features and linguistic features (language and discourse) to improve recognition performance. Typical examples of linguistic-paralinguistic-fusion methods are those of Litman et al. [48] and Schuller et al. [68], who used spoken words and acoustic features, of Lee and Narayanan [46], who used prosodic features, spoken words and information of repetition, and of Bartliner et al. [7], who used Part-of-speech (POS), dialogue act (DA), repetitions, corrections, and syntactic-prosodic boundary to infer the emotion. Litman et al. [48] investigated the role of the context information (e.g. subject, gender and problem, turn-level features representing local and global aspects of the prior 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 [3]. Second how to extract semantic discourse information is more challenging. As

a result, most of these features have been extracted manually or directly from transcripts.

3.3 Audio-visual Expression Recognition

In the survey written by Pantic and Rothkrantz in 2003, [59], only four studies were found that were focused on audiovisual affect recognition. Since then, an increasing number of efforts are reported toward this direction. Although most of existing audio-visual affect recognition studies investigated recognition of basic emotions, fewer efforts are underway to detect non-basic emotion, i.e., those of Zeng et al. [85], [87], [88], who added 4 cognitive states (interest, puzzlement, frustration and boredom) considering the importance of these cognitive states in human computer interaction.

Recently a few studies have been reported toward audio-visual spontaneous emotion recognition [10], [30], [86]. These studies are that of Zeng et al. [86], who used the data collected in psychological research interview (Adult Attachment Interview), and of Fragopanagos and Taylor [30] and Caridakis et al. [10], who used the data collected in Wizard of OZ scenarios. Because their data were not sufficient to build classifiers for fine-grained affective states (e.g., basic emotions), they chose to recognize coarse affective states, e.g., positive and negative states in [86], or quadrants in evaluation-activation space [10], [30]. The studies [10], [30] applied the FeelTrace system that enables raters to continuously label the change of affective expressions. The study [30] noticed the considerable labeling variation among four raters using FeelTrace [19] due to subjectivity of audio-visual affect judgment. Specifically, one rater mainly relied on audio information to make judgment while another rater mainly relied on visual information. In order to reduce this variation, the studies [86] 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.

Three fusion strategies (feature-level, decision-level and model-level fusions) are found to be used in the audio-visual affect recognition. A typical example of feature-level fusion is the study [9], which concatenated the prosodic features and facial features to construct joint feature vectors that are then used to build an affect recognizer. However, the different time scale and metric level of features from different modalities and increasing feature dimension influence the performance of the feature-level fusion. Most of the bimodal affect recognition studies applied decision-level fusion (e.g., [9], [31], [38], [79], [88]), which independently model audio-only and visual-only expressions, then combine these uni-modal recognition results at the end. Since humans display audio and visual expressions in a complementary and redundant manner, the conditional independent assumption of decision-level fusion actually loses the correlation information between audio and visual signals. Some interesting model-level fusion methods are introduced that can make use of the correlation between audio and visual streams, and relax the requirement of synchronization of these streams. Zeng et al. [87] 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. [85] 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. [73] 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 [30] proposed an artificial neural network with a feedback loop called ANNA to integrate the information from face, prosody and lexical content. Caridakis et al. [10] investigated combining face and prosody expressions by using Relevant Neural Networks.

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 finally addressed. 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).

Here we focus on discussing the challenges in computing methods for developing of automatic spontaneous affect recognizer. As for the challenges to spontaneous emotion database collection and annotation, the readers are referred to [18], [21], [34], [59], [62].

4.1 Visual Input

Development of vision processing techniques that are robust in fully unconstrained environments is still in the relatively distant future. The existing visual face detection and tracking techniques are just able to reliably handle the near-front/profile view of face images with good resolution and lighting conditions. In a realistic interaction environment, the arbitrary movement of subjects, low-resolution and hand occlusion can cause these techniques to fail. The view-independent facial expression recognition based on 3D face model is worthy of further investigation [12], [83]. Development of a robust face detector, head and facial feature tracker forms the first step in the realization of facial expression analyzers capable of handling unconstrained environment.

In a realistic interaction environment, a facial expression analyzer should be able to deal with noisy and partial data and to generate its conclusion with confidence that reflects uncertainty of output of face and face point localization and tracking. Further efforts are needed toward modeling the static and dynamic structure of facial expression in order to handle noise features, temporal information, and partial data.

Except for few studies (e.g., [4], [16], [35], [40], [90]), the existing efforts analyzed facial expression behavior isolated from other visual cues (eye and head movement, and body gesture). It is suggested in the study [44] that multimodal coordination of facial expression, head movement and gesture is important to judge certain affect expression such as embarrassment. Integration of these multiple cues for automatic visual-based affect recognition is a largely unexplored research.

4.2 Audio Input

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) have been introduced for affect recognition in literature, the optimal feature set has not yet been established from the existing experiments.

Another challenge is how to reliably automatically extract these linguistic and paralinguistic feature from the audio channel. When we analyze the prosody in realistic conversation, we have to consider the multiple functions of prosody that include expression of affect and a variety of linguistic function [53]. Prosody features can be used to indicate discourse and segmentation information not only to express emotion. The prosodic event model that can reflect these functions simultaneously is worthy of further investigation. In addition, automatic extraction of spoken words from spontaneous emotional speech is also a difficult problem because the recognition rate of the exiting automatic speech recognition (ASR) system is far from perfect. The emotional aspects in speech further reduce ASR performance [3]. The automatic extraction of high-level underlying semantic linguistic information (e.g. dialogue act, repetitions, corrections, and syntactic information) is more challenging.

4.3 Fusion

Although the benefit of fusion (i.e., audio-visual fusion, linguistic and paralinguistic fusion, multi-visual-cue fusion from face, head and body gestures) for affect recognition is expected from engineering and psychological perspectives, our knowledge of how humans achieve this fusion is extremely limited. The neurological studies on fusion of sensory neurons [72] seem to more support early fusion (i.e., feature-level fusion) than late 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 dynamic structures, based on existing methods. Due to these difficulties, most researchers choose decision-level fusion that simplifies the fusion problem by introducing the conditional dependent assumption. Model-level fusion or hybrid fusion that combines the benefits of both feature-level and decision-level fusion methods may be the best choice for this fusion problem. Based on existing knowledge and methods, how to model multimodal fusion 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.

4.4 Context

Investigation is clearly warranted to address how to make use of contextual information to improve the performance of affect recognition. Emotions are intimately related to a situation being experienced or imagined by human. Without context, human may misunderstand speaker's emotion expressions. Since the problem of context sensing is very difficult to solve, pragmatic approaches (e.g. activity- and user-profiled approaches) should be taken when learning the grammar of human affective behavior [57]. Yet, with the exception for a few studies (e.g., [29], [40], [48], [60]), 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 field like face recognition, gender recognition, age recognition, topic detection, and task tracking.

4.5 Evaluation

Unfortunately, the diverse methods reviewed in this paper are difficult to compare because they are rarely tested on a common experimental condition (e.g., data and annotation). United efforts of different research communities are needed to address the evaluation of system performance based on a comprehensive, readily accessible benchmark database with annotation.

5. CONCLUSION

In the comprehensive survey written by Pantic and Rothkrantz in 2003 [59], almost all automatic affect recognition efforts were based small artificial emotion data, and only four studies were focused on audio-visual affect recognition. Since then, the picture has changed considerably. Increasing efforts are reported toward recognition of spontaneous affective expression by using audio and visual information and fusion methods. Some pilot studies have identified some problems that have been missed or avoided in uni-modal posed emotion recognition.

The shifts of perspective in affect recognition research, from uni-modal to multimodal and from posed emotion expression to spontaneous emotion expression, in turn highlight many challenges to our knowledge and existing techniques. Collaboration among related disciplines is certainly the most powerful means to advance our knowledge on the nature of affect, and in turn enhance automatic affect recognition performance.

6. Acknowledgment

The authors would like to thank reviewers for valuable comments on this paper. This paper is collaborative work. Thomas Huang is the lead of this team work but likes to be the last in the author list as usual. Zhihong Zeng wrote the first draft, Maja Pantic significantly improved it by rewriting it and offering important advices, and Glenn Roisman provided important comments and polished it. The work is supported in part by a Beckman Postdoctoral Fellowship and in part by National Science Foundation Grant CCF 04-26627.

7. REFERENCES

[1] 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

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

[3] 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

[4] 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

[5] 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

[6] 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
[7] Batliner A, Fischer K, Hubera R, Spilkera J and Noth E. (2003). How to find trouble in communication. Speech Communication, Vol. 40, 117-143.
[8] 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.
[9] 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
[10] 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
[11] Chang Y, Hu C, Turk, M (2004). Probabilistic expression analysis on manifolds. Proc. Computer Vision and Pattern Recognition, 2:520-527
[12] 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.
[13] 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
[14] 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
[15] Cohn, J.F. (2006), Foundations of Human Computing: Facial Expression and Emotion, Int. Conf. on Multimodal Interfaces, 233-238
[16] 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
[17] 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
[18] Cowie R, Douglas-Cowie E and Cox C (2005). Beyond emotion archetypes: databases for emotion modeling using neural networks. Neural Networks, 18: 371-388
[19] 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
[20] 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
[21] Devillers L, Vidrascu L, and Lamel L (2005). Challenges in real-life emotion annotation and machine learning based detection. Neural Networks, 18: 407-422
[22] 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
[23] Ekman P. (1972). Universals and cultural differences in facial expressions of emotion. Nebr. Symp. Motiv. 1971, 207-283
[24] Ekman, P., editor (1982). Emotion in the human face. Cambridge University Press, New York, 2nd edition
[25] Ekman, P., Friesen, W.V., Hager, J.C. (2002). Facial Action Coding System. A Human Face, Salt Lake City, USA
[26] Ekman P. and Oster H. (1979). Facial expressions of emotion. Ann. Rev. Psychol. 1979, 30:527-554
[27] 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.
[28] 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
[29] 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
[30] Fragopanagos, F. and Taylor, J.G. (2005), Emotion recognition in human-computer interaction, Neural Networks, 18: 389-405
[31] Go H.J, 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
[32] 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
[33] Grimm, M. and Kroschel, K. (2005). Evaluation of Natural Emotions Using Self Assessment Manikins, IEEE Workshop on Automatic Speech Recognition and Understanding, 381-383
[34] Gross, R. (2005).Face databases. In: Handbook of Face Recognition, Li S Z., Jain A.K., (Eds.), Springer, New York, USA, 301-328
[35] 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
[36] 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
[37] Hirschberg, J., Benus, S., Brenier, J.M., Enos, F., Friedman, S. (2005). Distinguishing Deceptive from Non-Deceptive Speech. Interspeech, 1833-1836
[38] Hoch, S., Althoff, F., McGlaun, G., Rigoll, G. (2005), Bimodal fusion of emotional data in an automotive environment, ICASSP, Vol. II, 1085-1088, 2005
[39] 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.
[40] 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
[41] 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
[42] 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
[43] Kapoor, A., Burleson, W., and Picard, R. W. (2007), Automatic prediction of frustration. Int. Journal of Human-Computer Studies. Vol. 65(8), 724-736.
[44] 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
[45] Kwon, O.W., Chan, K., Hao, J., Lee, T.W (2003), Emotion Recognition by Speech Signals, EUROSPEECH.
[46] Lee C M Narayanan, S.S. (2005). Toward detecting emotions in spoken dialogs. IEEE Tran. Speech and Audio Processing, Vol. 13(2): 293-303
[47] Liscombe, J., Hirschberg, J., Venditti, J.J. (2005). Detecting Certainness in Spoken Tutorial Dialogues. Interspeech.
[48] 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
[49] Littlewort G, Bartlett M S and Lee K (2006). Faces of Pain: Automated measurement of spontaneous facial expressions of

genuine and posed pain, 13 Joint Symposium on Neural Computation. 1
[50] 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
[51] Maat, L., Pantic, M. (2006). Gaze-X: Adaptive affective multimodal interface for single-user office scenarios, Proc. ACM Int'l Conf. Multimodal Interfaces, 171-178
[52] Neiberg D, Elenius K, and Laskowski K. (2006). Emotion Recognition in Spontaneous Speech Using GMM. Int. Conf. on Spoken Language Processing, 809-812
[53] Mozziconacci, S. (2002). Prosody and Emotions. Int. Conf. on Speech Prosody.
[54] 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.
[55] Pantic, M., and Bartlett, M.S. (2007). Machine analysis of facial expres-sions. In Face Recognition, Delac, K. and Grgic, M., Eds. Vienna, Austria: I-Tech Education and Publishing, 377-416
[56] Pantic M and Patras T (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
[57] 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
[58] 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
[59] Pantic M., Rothkrantz, L.J.M. (2003), Toward an affect-sensitive multimodal human-computer interaction, Proceedings of the IEEE, Vol. 91, No. 9, Sept., 1370-1390
[60] 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
[61] Pantic, M., Sebe, N., Cohn, J.F., and Huang, T.S. (2005), Affective Multimodal Human-Computer Interaction, ACM Multimedia, 669-676
[62] 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
[63] Pentland, A. (2005). Socially aware, computation and communication, IEEE Computer, Vol.38, 33-40
[64] Picard, R.W. (1997). Affective Computing, MIT Press, Cambridge.
[65] Plutchik R. (1980). Emotion: A psychoevolutionary synthesis. New York: Harper and Row.
[66] 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
[67] Russell J.A., Bachorowski J. and Fernandez-Dols J. (2003). Facial and vocal expressions of emotion. Ann. Rev. Psychol. 54:329-349
[68] 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
[69] Sebe, N., Cohen, I., and Huang, T.S. (2005). Multimodal Emotion Recognition, Handbook of Pattern Recognition and Computer Vision, World Scientific, 2005.
[70] 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
[71] 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
[72] Stein, B., Meredith, M.A. (1993). The Merging of Senses. MIT Press, Cambridge, USA
[73] 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
[74] 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,
[75] 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
[76] Truong K P and van Leeuwen D A (2005). Automatic detection of laughter. In Proc. Interspeech, pp. 485-488, 2005.
[77] 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
[78] Valstar MF, Pantic M, Ambadar Z and Cohn JF. (2006). Spontaneous vs. Posed Facial Behavior: Automatic Analysis of Brow Actions. Int. Conf. on Multimedia Interfaces. 162-170
[79] Wang, Y. and Guan, L.(2005), Recognizing human emotion from audiovisual information, ICASSP, Vol. II, 1125-1128
[80] 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
[81] 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
[82] 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
[83] 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
[84] 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.
[85] 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
[86] Zeng Z, Hu Y, Roisman G I, Wen Z, Fu Y and Huang T S (2006): Audio-visual emotion recognition in adult attachment interview. Int. Conf. Multimodal Interfaces: 139-145
[87] 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
[88] 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
[89] Zhang T, Hasegawa-Johnson M and Levinson S E (2004). Children's Emotion Recognition in an Intelligent Tutoring Scenario, Interspeech 2004.
[90] 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