SIP (2014), vol. 3, e12, page 1 of 18 © The Authors, 2014.

doi:10.1017/ATSIP.2014.11

OVERVIEW PAPER

Survey on audiovisual emotion recognition: databases, features, and data fusion strategies

CHUNG-HSIEN WU[1], JEN-CHUN LIN[1,2] AND WEN-LI WEI[1]

Emotion recognition is the ability to identify what people would think someone is feeling from moment to moment and understand the connection between his/her feelings and expressions. In today's world, human–computer interaction (HCI) interface undoubtedly plays an important role in our daily life. Toward harmonious HCI interface, automated analysis and recognition of human emotion has attracted increasing attention from the researchers in multidisciplinary research fields. In this paper, a survey on the theoretical and practical work offering new and broad views of the latest research in emotion recognition from bimodal information including facial and vocal expressions is provided. First, the currently available audiovisual emotion databases are described. Facial and vocal features and audiovisual bimodal data fusion methods for emotion recognition are then surveyed and discussed. Specifically, this survey also covers the recent emotion challenges in several conferences. Conclusions outline and address some of the existing emotion recognition issues.

Keywords: Emotion recognition, human-computer interaction, audiovisual data fusion

Received 12 November 2013; Revised 13 October 2014

I. INTRODUCTION

Emotion plays an important role in social interaction, human intelligence, perception, etc. [1]. Since perception and experience of emotion are vital for communication in the social environment, understanding emotions becomes indispensable for the day-to-day function of humans. Technologies for processing daily activities, including facial expression, speech, and language have expanded the interaction modalities between humans and computer-supported communicational artifacts, such as robots, iPad, and mobile phones. With the growing and varied uses of human–computer interactions, emotion recognition technologies provide an opportunity to promote harmonious interactions or communication between computers and humans [2, 3].

Basically, emotion could be expressed through several social behaviors, including facial expression, speech, text, gesture, etc. According to human judgment of affect, psychologists have various opinions about the importance of the cues from facial expression, vocal expression and linguistic message. Mehrabian stated that the facial expression of a message contribute 55% of the overall impression while the vocal part and the semantic contents contribute 38 and 7%, respectively [4]. Among these modalities, facial expression is acknowledged as one of the most direct channels to transmit human emotions in non-verbal communication [5, 6]. On the one hand, speech is another important and natural channel to transmit human affective states especially in verbal communication. Affective information in speech can be transmitted through explicit (linguistic) and implicit (paralinguistic) messages during communication [7]. The former can be understood and extracted from affective words, phrases, sentences, semantic contents, and more. The later may be explored from prosodic and acoustic information of speech. In the past years, analysis and recognition approaches of artificial affective expressions from a uni-modal input have been widely investigated [8–11]. However, the performance of emotion recognition based on only facial or vocal modality still has its limitation. To further improve emotion recognition performance, a promising research area is to explore the data fusion strategy for effectively integrating facial and vocal cues [12–14]. In face-to-face communication, humans employ these communication paths alone or using one to complement and enhance another. But the roles of multiple modalities and their interplay remain to be quantified and scientifically understood. In the past 3 years, Audio/Visual Emotion Challenges (AVEC 2011–2013) [15–17] aimed to compare audiovisual signal processing and

[1] Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. Phone: +886 6 208 9349
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan

Corresponding author:
Chung-Hsien Wu
Email: chunghsienwu@gmail.com

machine learning methods to advance emotion recognition systems. Data fusion strategy effectively integrating the facial and vocal cues has become the most important issue.

This paper gives a survey on the existing audiovisual emotion databases and recent advances in the research on audiovisual bimodal data fusion strategies. This paper also introduces and surveys the recent emotion challenges which have been conducted in *ACM Multimedia*, *ICMI*, *ACII*, *INTERSPEECH*, and *FG*, specifically the AVEC 2011–2014 [15–18], 2009 Emotion challenge [19], 2010 Paralinguistic Challenge [20], 2013 COMputational PARalinguistics challengE (ComParE) [21], Facial Expression Recognition and Analysis (FERA) 2011 challenge [22] and Emotion recognition in the Wild (EmotiW) 2013 challenge [23].

The rest of the paper is organized as follows. Section II describes the existing audio-visual emotion databases. Section III presents the state-of-the-art audiovisual bimodal data fusion strategies and introduces the utilized audio/facial features and classifiers. Finally, Section IV offers the conclusion.

II. AUDIOVISUAL EMOTION DATABASES

Audiovisual emotion databases play a key role in emotion recognition for model training and evaluation. Numerous achievements have been reported on the collection of the emotion databases [12, 24, 25]. For instance, the Association for the Advancement of Affective Computing (AAAC), formerly the HUMAINE Association, provided several multimodal, speech, and facial expression databases for association members. Although rich emotion databases were collected for different application purposes, most of the databases were constructed between 1996 and 2005. Moreover, the modalities of the multimodal emotion databases include not only speech and facial expressions but also texts, bio-signals, and body poses. In this paper, we focused only on emotion databases for audiovisual processing. This paper also provided the detailed characteristics of the currently available benchmark databases between 2006 and 2014 which were commonly used in audiovisual emotion recognition studies and emotion challenges from facial and vocal expressions. Table 1 summarizes the databases and some of the noteworthy data resources for audiovisual emotion recognition task. This table describes the following information about each database:

(1) Name of database,
(2) Language of recordings,
(3) Affective state elicitation method (posed (acted), induced or spontaneous emotional expression) [12, 25–27],
(4) Number of subjects,
(5) Number of available data samples,
(6) Affective state description (category, dimension or event) [12, 25, 28],
(7) Availability,
(8) Publication year, and
(9) Challenges used or references reported.

A) Elicitation method

The affective state elicitation methods for the collection of the audiovisual databases reported in the literature can be classified into three major categories: (a) posed (acted), (b) induced (via clips), and (c) spontaneous (occurring during an interaction) [12, 25–27]. With regard to the posed databases [29, 47], every actor was asked to correctly express each emotion. For instance, the GEneva Multimodal Emotion Portrayal (GEMEP) database [29, 58, 59] consists of more than 7000 audio-video emotion portrayals, which were portrayed by ten professional actors with the help of a professional theater director. The GEMEP was selected as the database for the emotion sub-challenge of the INTERSPEECH 2013 ComParE [21]. A subset of the GEMEP database was also used as the dataset for facial Action Unit (AU) recognition sub-challenge of the FERA 2011 challenge [22].

In terms of the induced databases [30], a subject's emotional responses are commonly evoked by films, stories, music, etc. For example, the eNTERFACE'05 EMOTION Database [30, 60] was designed and collected during the eNTERFACE'05 workshop. Each subject was asked to listen to six successive short stories, each eliciting a particular emotion. If two human experts judged the reaction expressing the emotion in an unambiguous way, then the sample was added to the database. However, in recent years the study of emotion recognition on the expressed stances has gradually moved from posed or induced expressions to more spontaneous expressions.

According to previous studies, the important issues for natural emotion database collection include spontaneous emotion and conversational elements. The audiovisual data with spontaneous emotion are difficult to collect because the emotion expressions are relatively rare, short-lived, and often associated with a complex contextual structure. In addition, the recorded data in most emotion databases were not produced in a conversational context, which limits the naturalness of temporal course of emotional expressions, and ignores the response to different situations. To deal with these problems, some of the existing databases were collected based on interactive scenarios including human–human (dyadic) conversation [33, 44, 57] and human–computer interaction (HCI) [17, 46]. To this end, the Sensitive Artificial Listeners (SAL) scenario [61] was developed from the ELIZA concept introduced by Weizenbaum [62]. The SAL scenario can be used to build a spontaneous emotion database through machine agent–human conversation, which tries to elicit the subject's emotions. However, in previous SAL recordings, an operator navigated a complex script and decided what the "agent" should say next. These recordings were emotionally interesting, but the operator conducted a conversation with quite minimal understanding of the speech

Table 1. Audiovisual databases for emotion recognition task.

Database	Language	Elicitation method	# of subjects	# of samples	Emotion description	Available	Year	Challenge used/Ref. reported
GEMEP [29]	French	Posed (portrayed by professional actors with the help of a professional theatre director)	Ten professional actors (five males, five females)	Over 7000 portrayals	18 affective states (5 discrete emotion classes: anger, fear, joy, relief, sadness was used in the FERA 2011)	Yes	2006	FERA 2011; INTERSPEECH 2013 ComParE
eNTERFACE '05 [30]	English	Induced (elicited from listen a short story)	42 subjects (34 man, 8 woman from 14 different nationalities)	1166 video sequences	Six emotion categories (anger, disgust, fear, happiness, sadness, surprise)	Yes	2006	eNTERFACE '05 workshop; [31, 32]
IEMOCAP [33]	English	Acted, Spontaneous (affective dyadic interaction with markers on the face, head, and hands) (both improvised and scripted sessions)	Ten actors (five males, five females)	12 h	Five emotion categories (happiness, anger, sadness, frustration, and neutral); 3 dimensions (valence, activation, dominance)	Yes	2007	[34–37]
RML [38]	six languages	Acted	Eight subjects	500 video samples	Six emotion categories (anger, disgust, fear, happiness, sadness, surprise)	Yes	2008	[39]
VAM [40]	German	Spontaneous (TV talk-show)	47 talk show guests	947 utterances (approximately 12 h)	Three dimensions (valence (negative vs. positive), activation (passive vs. active), dominance (weak vs. strong))	Yes	2008	[41]
SAVEE [42]	English	Acted	Four male actors	480 utterances	Seven emotion categories (anger, disgust, fear, happiness, sadness, surprise, neutral)	Yes	2009	[43]
TUM AVIC [44]	English	Spontaneous (natural human-to-human conversational speech of a product presentation)	21 subjects	3901 turns	Five level of interest; 5 non-linguistic vocalizations (breathing, consent, garbage, hesitation, laughter)	Yes	2007	INTERSPEECH 2010 Paralinguistic Challenge; [45]
SEMAINE [46]	English	Spontaneous (conversations between humans and artificially intelligent agents)	150 participants	959 conversations (24 recordings for the AVEC challenge)	27 associated categories; 5 affective dimensions (valence, activation, power, expectation, overall emotional intensity)	Yes	2010	AVEC 2011; 2012; 2013; [15, 16, 37, 47–50]

(*continued*)

Table 1. Continued.

Database	Language	Elicitation method	# of subjects	# of samples	Emotion description	Available	Year	Challenge used/Ref. reported
MHMC [47]	Chinese	Posed (actor must ensure that the particular emotion is properly vocalized and expressed)	7 actors (both genders)	1680 Sentences (approximately 5 h)	Four emotion categories (happiness, sadness, anger, neutral)	Upon request	2011	[47, 50, 51]
AFEW [52]	English	Spontaneous (extracted from movies in the wild)	330 subjects (single and multiple subjects per sample, age range from 1 to 70 years)	1426 sequences	Seven emotion categories (anger, disgust, fear, happiness, neutral, sadness, surprise)	Yes	2012	EmotiW
Spanish Multimodal Opinion [53]	Spanish, English	Spontaneous (collected from the social media web site YouTube)	105 speakers	105 videos	Positive, negative	Upon request	2013	[54]
MAHNOB Laughter [55]	Mother language, English	Spontaneous, Posed (first session: recorded while watching funny video clips; second and third sessions: pose a smile and produce an acted laughter, respectively)	22 subjects (12 males, 10 females)	180 sessions (a total duration of 3 h 49 m)	Laughter, speech, posed laughter, speech laughter, other vocalizations	Yes	2013	[56]
AVDLC [17]	German, English	Spontaneous (HCI task)	292 subjects (age range from 18 to 63 years)	340 video clips	Minimal depression, mild depression, moderate depression, severe depression	Yes	2013	AVEC 2013; 2014
RECOLA [57]	French	Spontaneous (remote dyadic collaborative interactions)	46 subjects (19 males, 27 females)	7 h	Five social behaviors (agreement, dominance, engagement, performance, rapport); arousal and valence	Yes	2013	[24]

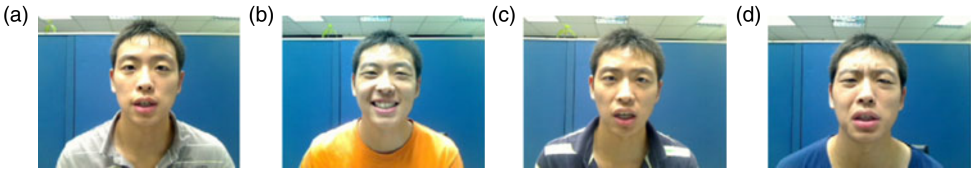


Fig. 1. Examples for the posed expression with four emotional states: (a) Neutral, (b) Happy, (c) Angry, and (d) Sad.

content. Therefore, the conversational elements were very limited. Then the "Solid SAL" scenario was recently developed to overcome this problem; that is, the operator was requested to be thoroughly familiar with the SAL characters, and spoke as they would without navigating a complex script. The most famous example is the public benchmark database "SEMAINE" [46, 63, 64]. For data recording of the SEMAINE database, the participant selected one of four characters (i.e. Prudence, Poppy, Spike, and Obadiah) of the operator to interact with. In AVEC 2011 and 2012 [15, 16], part of the SEMAINE database was used as the benchmark database.

As mentioned above, toward robust automatic emotion recognition, collecting the data with spontaneous emotion and conversational element is valuable and important. However, current audiovisual expression databases have been recorded in laboratory conditions lacking available data with real-world or close-to-real-world conditions. Accordingly, in EmotiW 2013 challenge [23], the Acted Facial Expressions in the Wild (AFEW) database was collected from movies and formed the bases providing a platform for researchers to create, extend and verify their methods on real-world data. The AFEW database was collected by searching the closed caption keywords (e.g. [HAPPY], [SAD], [SURPRISED], [SHOUTS], [CRIES], [GROANS], [CHEERS], etc.) which were then validated by human annotators. In addition, Spanish Multimodal Opinion database [53] also collected a dataset consisting of 105 videos in Spanish from YouTube. The videos were found using the following keywords: *mi opinion* (my opinion), *mis producto favoritos* (my favorite products), *me gusta* (I like), *no me gusta* (I dislike), *producto para bebe* (baby products), *mis perfumes favoritos* (my favorite perfumes), *peliculas recomendadas* (recommended movies), *opinion politica* (politic opinion), *video juegos* (video games), and *abuso animal* (animal abuse).

B) Emotion categorization

In light of emotion characterization in the databases, emotions are categorized into three representations for emotion recognition: (a) discrete categorical representation, (b) continuous dimensional representation, and (c) event representation (affective behavior; e.g. level of interest, depression, laughter, etc.) [12, 25, 28]. Many of the emotion recognition studies have attempted to recognize a small set of prototypical emotional states such as six prototypical emotions: anger, disgust, fear, happiness, sadness, and surprise proposed by Ekman [2, 65, 66]. Even though automatic facial/speech emotion recognition has been well studied, prototypical emotions cover only a subset of the total possible facial/speech expressions. For example, boredom, and interest cannot seem to fit well in any of the prototypical emotions. Moreover, the collection of audio or visual emotional signals in some categories such as fear or disgust is not easy. Accordingly, several studies [34, 36, 47, 50, 51] focused on recognition of more familiar emotion categories such as happy, sad, angry, and neutral which were comparatively easy to express. For example, the recognition work using the posed MHMC database [47] focused on these four emotional categories. Figure 1 shows some example images for the four emotional states in the MHMC database. However, these emotions only represent a small set of human affective states, and are unable to capture the subtle affective change that humans exhibit in everyday interactions.

To accommodate such subtle affective expressions, researchers have begun adopting a dimensional description of human emotion where an emotional state is characterized in numerous latent dimensions [67, 68]. Examples of the affective dimensions, such as Activation/Arousal (passive/active), Expectation (anticipation), Power/Dominance (sense of control, i.e. weak/strong), and Valence (negative/positive), have been well established in the psychological literature. The problem of dimensional emotion recognition can thus be posed as a regression problem or reduced into a binary-class classification problem [28] (active versus passive, positive vs. negative, etc.) or even as a four-class classification problem (classification into quadrants of an arousal-valence two-dimensional (2D) space as shown in Fig. 2 [69–71]). An example of this category is the Vera-Am-Mittag (VAM) German audiovisual spontaneous database [40]. The VAM database consists of audio-visual recordings taken from a German TV talk show and was annotated along three emotional dimensions: valence, arousal, and dominance. In order to label the dimensions in continuous time and continuous value in AVEC 2011 and 2012 challenges, a tool called FeelTrace [72] was used to annotate the SEMAINE database.

In addition, in the INTERSPEECH 2010 Paralinguistic Challenge, AVEC 2013, and 2014, and LREC Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES^3) 2012, event recognition has also been an object of study. Since emotion, social signals, and sentiment from text are part of social communication, recognition of events, including signals such as laugh, smile, sigh, hesitation, consent, etc. are highly relevant in helping better understand affective behavior and its context. For

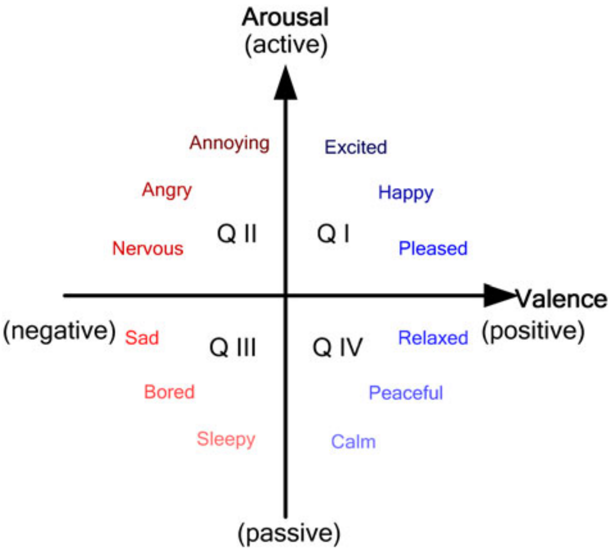


Fig. 2. Valence-activation 2D emotion plane [69, 70].

instance, understanding a subject's personality is needed to make better sense of observed emotional patterns and non-linguistic behavior, that is, laughter and depression analysis can give further insight into the personality trait of the subject. Different from discrete categorical and continuous dimensional emotion prediction, event-based recognition, such as level of depression, has given us new opportunities and challenges. In the INTERSPEECH 2010 Paralinguistic Challenge [20], audio part of TUM Audio-Visual Interest Corpus (TUM AVIC) [44, 73] was used and labeled in three levels of interest from boredom (level of interest 1 (loi1)), over neutral (loi2) to joyful (loi3) interaction. In AVEC 2013 and 2014 [17, 18, 74], a subset of Audio-Visual Depressive Language Corpus (AVDLC) labeled with the level of self-reported depression was used for the depression recognition subchallenge. On the one hand, since both laughter and speech are also naturally audiovisual events, the MAHNOB Laughter audiovisual database [55] containing laughter, speech, posed laughs, speech-laughs, and other vocalizations was also created.

III. AUDIOVISUAL BIMODAL FUSION FOR EMOTION RECOGNITION

As it is difficult to include all of these studies, this paper introduces and surveys these advances for the recent research on audiovisual bimodal data fusion strategies for emotion recognition. Table 2 lists the existing popular data fusion strategies for facial–vocal expression-based emotion recognition with respect to the utilized database, type of emotion categorization, audio, and facial features, recognition methods (i.e. audio classifier (A), visual classifier (V), and audiovisual bimodal fusion approach (AV)), classifier fusion modality, recognition performance, and publication year.

A) Audio features

An important issue for emotion recognition from speech is the selection of salient features. Numerous features such as prosodic and acoustic features of emotional speech signals have been discussed over the years [78–82]. Among these features, prosodic features have been found to represent the most significant characteristics of emotional content in verbal communication and were widely and successfully used for speech emotion recognition [12, 83, 84]. Several studies have shown that pitch- and energy-related features are useful to determine emotion in speech [12, 47, 50, 85]. Morrison *et al.* [80] further summarized the correlations between prosodic features and emotions as shown in Table 3. In this survey, feature statistics in the MHMC database [50] are explored and used to illustrate the difference of the energy and pitch features among four emotional states (happy, angry, sad, and neutral). The distributions of the energy and pitch values for the four emotional states are shown in Fig. 3. According to our observations from the mean value of the energy feature, happy, and angry emotional states have higher intensities compared to sad and neutral states as shown in Fig. 3(a). In pitch features, the pitch levels and pitch ranges of sad emotion are lower and narrower than those of other emotional states; the mean and standard deviation of pitch in sad emotion are smaller than those of other emotions in Fig. 3(b). For energy feature, an ANOVA test [86] was applied to test the difference of the extracted energy values from speech frames among the four emotions. The ANOVA test results show that the difference of energy feature among the four emotional states is statistically significant ($F(3, 43\,977) = 14\,196.709$, $p < 0.0001$). Similarly, the ANOVA test was also applied to pitch features and demonstrated the statistical significance ($F(3, 43\,977) = 13\,173.271$, $p < 0.0001$). The results indicate that using pitch and energy is beneficial to emotion recognition for the posed MHMC database [50]. Even though these properties are obtained from the posed database, the findings are somewhat in accordance with the results of the previous studies based on natural emotion databases [80].

Besides energy and pitch features, voice quality features such as Harmonics-to-Noise Ratio (HNR), jitter, or shimmer, and spectral and cepstral features such as formants and Mel-Frequency Cepstral Coefficients (MFCCs) were also frequently used and discussed for emotion recognition [8, 9, 15–17, 19–21]. Referring to Table 2, Lin *et al.* [47, 51] and Wu *et al.* [50] used "Praat" [87] to extract three types of prosodic features, pitch, energy, and formants F1–F5 in each speech frame for emotion recognition. In [32, 34, 56], the MFCCs were used as audio features, which capture some local temporal characteristics. For instance, Metallinou *et al.* [34] used a 39D feature vector consisting of 12 MFCCs and energy, and their first and second derivations. In AVEC, 2011–2014 and 2009–2013 INTERSPEECH challenges, Schuller *et al.* utilized the open source software openSMILE [88] to extract Low-Level Descriptors (LLDs) features as the baseline feature set. The set of LLDs covers

Table 2. Literature review on facial–vocal expression-based emotion recognition.

Reference	Database	Class	Feature	Approach	Fusion modality	Result	Year
Schuller *et al.* [16]	SEMAINE	Arousal, Expectation, Power, Valence	(A) LLD/functional combinations (V) Local binary patterns	(A) SVR (V) SVR (AV) SVR	F	Average cross-correlation: (WLSC) (A) 0.027 (V) 0.011 (AV) 0.015	2012
Metallinou *et al.* [35]	IEMOCAP	Valence, Activation	(A) 12 MFCC coefficients, 27 Mel Frequency Bank (MFB) coefficients, pitch, energy, their first derivatives (V) The coordinates from 46 facial markers	(A) HMM (V) HMM (AV) BLSTM	F	Unweighted Accuracy: valence/activation (A) 49.99 ± 3.63/61.92 ± 4.88 (V) 60.98 ± 4.96/51.36 ± 4.14 (AV) 64.67 ± 6.48/52.28 ± 5.37	2012
Eyben *et al.* [45]	TUM AVIC	Garbage, Consent, Hesitation, Laughter	(A) 9 acoustic LLDs (V) 20 facial points	(A) LSTM-RNN (V) LSTM-RNN (AV) LSTM-RNN	F	Unweighted Average Recall (UAR) rate: (A) 67.6 (V) 41.1 (AV) 72.3	2012
Sayedelahl *et al.* [41]	VAM	Valence, Activation, Dominance	(A) Short-time energy, fundamental frequency, and 14 Mel frequency cepstral coefficients (V) Local binary patterns	(A) SVR with RBF kernel (V) SVR with RBF kernel (AV) SVR with RBF kernel	F	Average CC and (MLE) for the SPCA features: valence/activation/dominance (A) 0.62/0.80/0.79 (0.12/0.16/0.13) (V) 0.67/0.73/0.66 (0.11/0.18/0.16) (AV) 0.74/0.86/0.82 (0.09/0.13/0.12)	2013
Rosas *et al.* [53]	Spanish Multimodal Opinion	Positive, Negative	(A) Pause duration, pitch, intensity, loudness (V) Smile duration, gaze at camera	(A) SVM with linear kernel (V) SVM with linear kernel (AV) SVM with linear kernel	F	Accuracy (%): (A) 46.75 (V) 61.04 (AV) 66.23	2013
Rudovic *et al.* [56]	MAHNOB	Laughter, Speech	(A) 12 MFCCs (V) Feature points	(A) Logistic regression (V) Logistic regression (AV) Bimodal log-linear regression	F	Classification Rate (CR %): (A) 84.7 (V) 85.9 (AV) 92.7	2013
Metallinou *et al.* [34]	IEMOCAP	Anger, Happiness, Neutral, Sadness	(A) 39-dimensional MFCCs (V) The positions of facial markers are separated into six facial regions	(A) GMM (V) GMM (AV) Bayesian classifier weighting scheme	D	Classification accuracy (%): (A) 54.34 (V) 65.41 (AV) 69.59	2008
Metallinou *et al.* [36]	IEMOCAP	Anger, Happiness, Neutral, Sadness	(A) Mel filter bank coefficients (V) Facial marker coordinates	(A) HMM (V) GMM/HMM (AV) Bayesian fusion	D	Mean Unweighted accuracy (%UA): (A) 50.69 ± 5.14 (V) 55.74 ± 5.26 (AV) 62.27 ± 3.41	2010
Schuller *et al.* [15]	SEMAINE	Activity, Expectation, Power, Valence	(A) LLD/functional combinations (V) Local binary patterns	(A) SVMs with linear kernel (V) SVM with RBF kernel (AV) linear SVM	D	Mean Weighted Accuracy (%WA): (A) 45.1 (V) 46.2 (AV) 57.9	2011
Ramirez *et al.* [48]	SEMAINE	Activation, Expectancy, Power, Valence	(A) LLD/functional combinations (V) Horizontal and vertical eye gaze direction, smile intensity and head tilt	(A) LDCRF (V) LDCRF (AV) LDCRF	D	Average Weighted Accuracy (%WA): (A) 43.0 (V) 61.0 (AV) 60.3	2011

(*continued*)

Table 2. Continued

Reference	Database	Class	Feature	Approach	Fusion modality	Result	Year
Wöllmer *et al.* [49]	SEMAINE	Arousal, Expectation, Power, Valence	(A) LLD/functional combinations (V) Facial movement features	(A) BLSTM (V) SVM (AV) BLSTM	D	Mean Weighted Accuracy (%WA): (A) 65.2 (V) 59.3 (AV) 64.6	2013
Song *et al.* [75]	Self	Surprise, Joy, Anger, Fear, Sadness, Neutral	(A) 48 prosodic, 16 formant frequency features (V) Facial Animation Parameters (FAPs)	(A) HMM (V) HMM (AV) Tripled HMM (T-HMM)	M	Average Recognition Rate (%): (A) 81.08 (V) 87.39 (AV) 93.24	2008
Zeng *et al.* [76]	Self	4 cognitive states and 7 prototypical emotions	(A) Pitch, energy (V) 12 facial motion units	(A) HMM (V) HMM (AV) Multistream Fused HMM (MFHMM)	M	Average Accuracy: (A) 0.57 (pitch)/0.66 (energy) (V) 0.39 (AV) 0.80	2008
Paleari *et al.* [31]	eNTERFACE'05	Anger, Disgust, Fear, Happiness, Sadness, Surprise	(A) F0, first five formants, intensity, harmonicity, ten MFCC and 10 LPC (V) Facial FP absolute movements and relative movements of couples of facial FP	(A), (V) NN/SVM (AV) Neural Network based on Evidence Theory (NNET)	M	Mean Average Precision (MAP): (A) 0.253 (V) 0.211 (AV) 0.337	2009
Jiang *et al.* [32]	eNTERFACE'05	Anger, Disgust, Fear, Happiness, Sadness, Surprise	(A) 42-dimension MFCC (V) 18 facial features, 7 FAU	(A) HMM (V) HMM (AV) T_AsyDBN	M	Correction rates (%) (A) 52.19 (V) 46.78 (AV) 66.54	2011
Lin *et al.* [51]	MHMC	Neutral, Happy, Angry, Sad	(A) Pitch, energy, formants F1-F5 (V) 68 facial feature points from five facial regions	(A) HMM (V) HMM (AV) SC-HMM	M	Average recognition rate (%): (A) 67.75 (V) 67.25 (AV) 85.73	2011
Lu *et al.* [77]	Self	Valence, Activation	(A) Pitch F0, energy, and twelve MFCC features (V) 10 geometric distance features	(A) HMM (V) HMM (AV) Boosted Coupled HMM	M	Average recognition accuracies (%): valence/activation (A) 74.1/77.9 (V) 65.0/59.3 (AV) 92.0/90.2	2012
Wu *et al.* [50]	• MHMC • SEMAINE	• Happy, Sad, Angry, Neutral. Emotion quadrant I, II, III, IV	(A) Pitch, energy, formants F1-F5 (V) 30 FAPs	(A) HMM (V) HMM (AV) 2H-SC-HMM	M	Recognition rate (%): MHMC/SEMAINE (A) 71.01/60.31 (V) 71.37/62.19 (AV) 91.55/87.50	2013
Lin *et al.* [47]	• MHMC • SEMAINE	• Happy, Sad, Angry, Neutral. Emotion quadrant I, II, III, IV	(A) Pitch, energy, formants F1-F5 (V) 30 FAPs	(A) HMM (V) HMM (AV) EWSC-HMM	H	Recognition rate (%): MHMC/SEMAINE (A) 71.01/60.31 (V) 71.37/62.19 (AV) 90.59/78.13	2012

Feature: **LLD**: Low-Level Descriptors, **MFCC**: Mel-Frequency Cepstral Coefficients, **FAPs**: Facial Animation Parameters, **LPC**: Linear Predictive Coefficients, **FP**: Feature Points, **FAU**: Facial Animation Unit.
Approach: **SVR**: Support Vector Machine for regression, **LSTM-RNN**: Long Short-Term Memory Recurrent Neural Networks, **HMM**: Hidden Markov Model, **BLSTM**: Bidirectional Long Short-Term Memory neural network, **SVM**: Support Vector Machine, **GMM**: Gaussian Mixture Model.
LDCRF: Latent-Dynamic Conditional Random Field, **NN**: Neural Network, **T_AsyDBN**: Triple stream Asynchronous Dynamic Bayesian Network.
SC-HMM: Semi-Coupled HMM, **2H-SC-HMM**: Two-level Hierarchical alignment-based SC-HMM, **EWSC-HMM**: Error Weighted SC-HMM.
Fusion Modality: **F**eature/**D**ecision/**M**odel/**H**ybrid-level.
Result: **WLSC**: Word-Level Sub-Challenge, **CC**: Correlation Coefficient, **MLE**: Mean Linear Error, **SPCA**: Supervised Principal Component Analysis.

Table 3. Correlations among prosodic features and emotions [80].

	Pitch mean	Pitch range	Energy	Speaking rate	Formants
Anger	Increased	Wider	Increased	High	F1 mean increased; F2 mean higher or lower; F3 mean higher
Happiness	Increased	Wider	Increased	High	F1 mean decreased and bandwidth increased
Sadness	Decreased	Narrower	Decreased	Low	F1 mean increased and bandwidth decreased; F2 mean lower
Surprise	Normal or increased	Wider	–	Normal	–
Disgust	Decreased	Wider or narrower	Decreased or normal	Higher	F1 mean increased and bandwidth decreased; F2 mean lower
Fear	Increased or decreased	Wider or narrower	Normal	Higher or low	F1 mean increased and bandwidth decreased; F2 mean lower

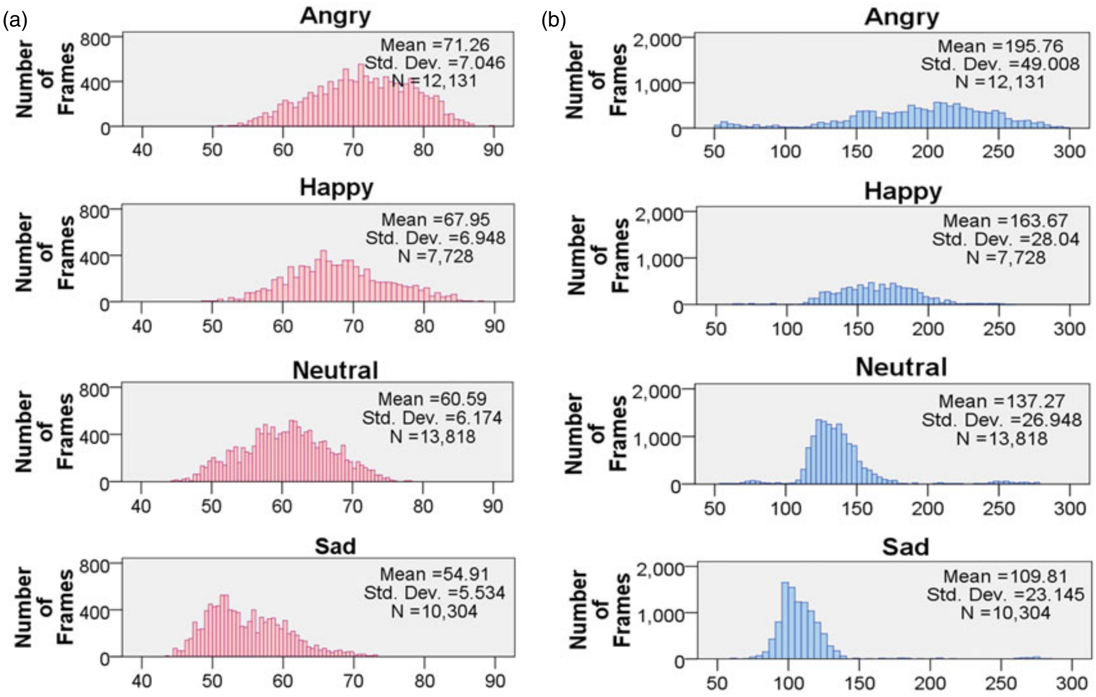
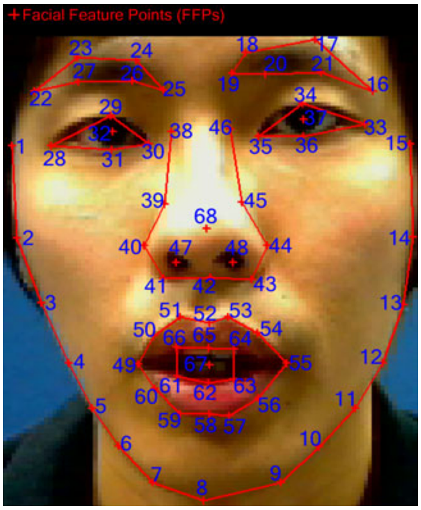


Fig. 3. The distributions of (a) energy and (b) pitch (Hz) for four emotional states in the posed MHMC database; N denotes the total number of frames.

a standard range of commonly used features in audio signal analysis and emotion recognition. For example, in INTERSPEECH 2009 Emotion challenge [19], the 16 LLDs are zero crossing rate (ZCR) from time signal, root-mean-squared (RMS) frame energy, pitch frequency, HNR by autocorrelation function, and MFCCs 1–12. The corresponding delta coefficients of the above features were also considered. Then 12 functionals, consisting of mean, standard deviation, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear regression coefficients with their mean-squared error (MSE), were applied. In addition, openEAR [89] has been widely used as an affect and emotion recognition toolkit for audio and speech affect recognition [16, 53].

For speech emotion recognition, speech features can be classified into two major categories including local (frame-level) and global (utterance-level) features according to the model properties [8, 90]. The local features represent the speech features extracted based on the unit of speech "frame". On the one hand, the global features are calculated from the statistics of all speech features extracted from the entire "utterance" [8]. For example, local features include spectral LLDs (e.g. MFCCs and Mel Filter Bank (MFB)), energy LLDs (e.g. loudness, energy), and voice LLDs (e.g. F0, jitter and shimmer); global features include the set of functionals extracted from the LLDs, such as max, min, mean, standard deviation, duration, linear predictive coefficients (LPC) [91]. Based on the extracted speech features (i.e. local or global features), traditional pattern recognition engines such as Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), support vector machine (SVM), etc. have been used in speech emotion recognition systems to decide the underlying emotion of the speech utterance. For instance, the dynamic modeling approach (e.g.

Table 4. The example of 68 facial feature points extracted using AAM alignment and related facial animation parameters.

Extracted facial feature points (FFPs)	Facial regions	FAPs Num.	Euclidean distance between FFPs	Comparing FFPs displacement with neutral frame
	Eyebrows	1, 2	$D_{\mathrm{vertical},1}(22, 30)$, $D_{\mathrm{vertical},2}(16, 35)$	$D_{v,1\_\mathrm{Neutral}}$-$D_{v,1}$, $D_{v,2\_\mathrm{Neutral}}$-$D_{v,2}$
		3, 4	$D_{\mathrm{vertical},3}(25, 30)$, $D_{\mathrm{vertical},4}(19, 35)$	$D_{v,3\_\mathrm{Neutral}}$-$D_{v,3}$, $D_{v,4\_\mathrm{Neutral}}$-$D_{v,4}$
		5, 6	$D_{\mathrm{vertical},5}(22, 28)$, $D_{\mathrm{vertical},6}(16, 33)$	$D_{v,5\_\mathrm{Neutral}}$-$D_{v,5}$, $D_{v,6\_\mathrm{Neutral}}$-$D_{v,6}$
		7, 8	$D_{\mathrm{vertical},7}(23, 28)$, $D_{\mathrm{vertical},8}(17, 33)$	$D_{v,7\_\mathrm{Neutral}}$-$D_{v,7}$, $D_{v,8\_\mathrm{Neutral}}$-$D_{v,8}$
		9, 10	$D_{\mathrm{vertical},9}(25, 28)$, $D_{\mathrm{vertical},10}(19, 33)$	$D_{v,9\_\mathrm{Neutral}}$-$D_{v,9}$, $D_{v,10\_\mathrm{Neutral}}$-$D_{v,10}$
		11, 12	$D_{\mathrm{vertical},11}(23, 30)$, $D_{\mathrm{vertical},12}(17, 35)$	$D_{v,11\_\mathrm{Neutral}}$-$D_{v,11}$, $D_{v,12\_\mathrm{Neutral}}$-$D_{v,12}$
		13	$D_{m,13}(19, 25)$	$D_{h,13\_\mathrm{Neutral}}$-$D_{h,13}$
	Eyes	14, 15	$D_{\mathrm{vertical},14}(29, 31)$, $D_{\mathrm{vertical},15}(34, 36)$	$D_{v,14\_\mathrm{Neutral}}$-$D_{v,14}$, $D_{v,15\_\mathrm{Neutral}}$-$D_{v,15}$
		16, 17	$D_{\mathrm{vertical},16}(28, 49)$,$D_{\mathrm{vertical},17}(33, 55)$	$D_{v,16\_\mathrm{Neutral}}$-$D_{v,16}$, $D_{v,17\_\mathrm{Neutral}}$-$D_{v,17}$
		18, 19	$D_{\mathrm{horizontal},18}(28, 30)$,$D_{\mathrm{horizontal},19}(33, 35)$	$D_{h,1\_18\mathrm{Neutral}}$-$D_{h,18}$, $D_{h,19\_\mathrm{Neutral}}$-$D_{h,19}$
	Nose	20, 21	$D_{\mathrm{vertical},20}(52, 68)$, $D_{\mathrm{vertical},21}(58, 68)$	$D_{v,20\_\mathrm{Neutral}}$-$D_{v,20}$, $D_{v,21\_\mathrm{Neutral}}$-$D_{v,21}$
		22, 23	$D_{\mathrm{vertical},22}(49, 68)$, $D_{\mathrm{vertical},23}(55, 68)$	$D_{v,22\_\mathrm{Neutral}}$-$D_{v,22}$, $D_{v,23\_\mathrm{Neutral}}$-$D_{v,23}$
	Mouth	24, 25	$D_{\mathrm{vertical},24}(52, 58)$, $D_{\mathrm{horizontal},25}(49, 55)$	$D_{v,24\_\mathrm{Neutral}}$-$D_{v,24}$, $D_{h,25\_\mathrm{Neutral}}$-$D_{h,25}$
	Facial Contours	26, 27	$D_{\mathrm{horizontal},26}(5, 58)$, $D_{\mathrm{horizontal},27}(11, 58)$	$D_{h,26\_\mathrm{Neutral}}$-$D_{h,26}$, $D_{h,27\_\mathrm{Neutral}}$-$D_{h,27}$
		28, 29	$D_{\mathrm{horizontal},28}(2, 68)$, $D_{\mathrm{horizontal},29}(14, 68)$	$D_{h,28\_\mathrm{Neutral}}$-$D_{h,28}$, $D_{h,29\_\mathrm{Neutral}}$-$D_{h,29}$
		30	$D_{\mathrm{vertical},30}(8, 68)$	$D_{v,30\_\mathrm{Neutral}}$-$D_{v,30}$

HMM) was applied to capture the temporal characteristics of affective speech and the detailed feature fluctuations for local feature vectors; the static modeling approach (e.g. GMM) was employed as the classifier for the global features.

B) Facial features

The commonly used facial feature types can be divided into appearance and geometric features [10, 11]. The appearance features depict the facial texture such as wrinkles, bulges, and furrows. The geometric features represent the shape or location of facial components (e.g. eyebrows, eyes, mouth, etc.). As the studies listed in Table 2, the IEMOCAP database [33] contains detailed facial marker information. Metallinou *et al.* [35, 36] used the (x,y,z) coordinates of 46 facial markers as the facial features. To capture face movements in an input video, in [45, 56], the Patras-Pantic particle filtering tracking scheme [92] was used to track 20 facial points. The 20 facial points consist of the corners/extremities of the eyebrows (4 points), eyes (8 points), nose (3 points), mouth (4 points), and chin (1 point). In addition, the features of shape and location can be estimated based on the results of facial component alignments through the classical approach, active appearance model (AAM) [93]. From previous research, the AAM achieved successful human face alignment, even for the human faces having non-rigid deformations. In [47, 50], the AAM was thus employed to extract the 68 labeled facial feature points (FFPs) from five facial regions, including eyebrows, eyes, nose, mouth, and facial contours, as shown in Table 4. For the purpose of normalization among different people, the facial animation parameters (FAPs) were expressed in terms of FAP units; each FAP unit represents a fraction of a key distance on the face. Then 30 FAPs were estimated from the vertical and horizontal distances from 24 out of 68 extracted FFPs as shown in Table 4. For example, the inner raised eyebrow FAPs were calculated as the distances by projecting vertically from the inner eyebrow feature points to the inner eye corners feature points, that is, from points 25 and 19 to points 30 and 35, which are further compared to their corresponding distances in the neutral frame. Similarly, Song *et al.* [75] used a set of 56 tracked FFPs to compute the 18 FAPs based on the displacement from the initial neutral face configuration. Jiang *et al.* [32] chose seven facial animation units (FAUs) features including AUV6-eyes closed, AUV3-brow lower, AUV5-outer brow raiser, AUV0-upper lip raiser, AUV11-jaw drop, AUV2-lip stretcher, and AUV14-lip corner depressor. For each image frame, the concatenation of 18 2D facial features and 7 FAU features, together with their first-order derivatives, results in a 50D feature vector as the facial features.

Being the dense local appearance descriptors, local binary patterns (LBPs) have been used extensively for facial expression recognition in recent years [94, 95]. LBPs were also used as the baseline features for the recent challenges in AVEC 2011–2014 challenges [15–18] and FERA 2011 challenge [22]. For example, Schuller *et al.* [15, 16] used LBP appearance descriptors as the facial features. After face and eye detection, the resulting face region was normalized based on eye locations. Then face region was divided into small blocks to extract LBP histograms through eight neighbors using binary comparisons. Finally, the LBP features extracted from each sub-region were concatenated into a feature vector (histogram) to represent a facial image. Furthermore, Rosas *et al.* [53] and Ramirez *et al.* [48] processed each video sequence with the Omron OKAO Vision software library [96]. The software automatically extracted the higher-level facial features from a subset of communicative signals such as horizontal and vertical eye gaze

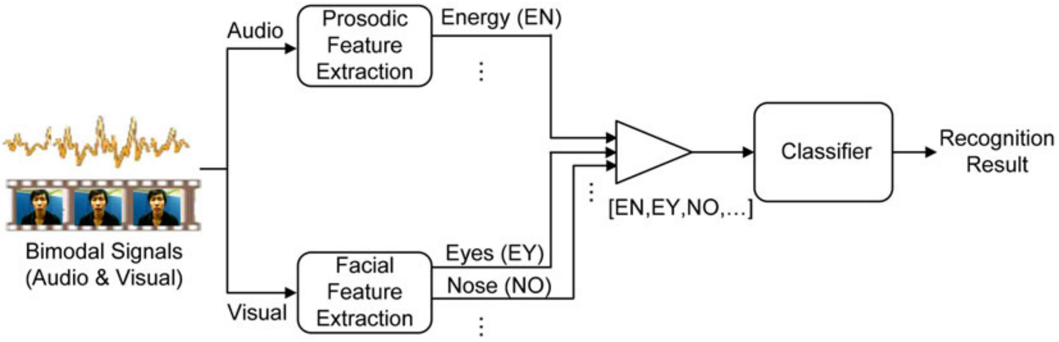

Audio
Prosodic Feature Extraction
Energy (EN)
Classifier
Recognition Result
[EN,EY,NO,...]
Bimodal Signals (Audio & Visual)
Facial Feature Extraction
Eyes (EY)
Nose (NO)
Visual

Fig. 4. Illustration of feature-level fusion strategy for audiovisual emotion recognition.

direction (degrees), smile intensity (from 0 to 100), and head tilt (degrees). This approach was shown to be useful when analyzing dyadic interactions for more naturalistic databases.

For audiovisual data fusion, to deal with the problem of mismatched frame rates between audio and visual features, the linear interpolation technique has been widely applied, which interpolates the video features to match the frame rate of audio features [32, 97]. In addition, some studies were based on reducing the frame rate of audio features in order to match the video features [47, 50].

C) Bimodal fusion approaches

Many data fusion strategies have been developed in recent years. The fusion operations in previous studies can be classified into feature-level (early) fusion, decision-level (late) fusion, model-level fusion, and hybrid approaches for audiovisual emotion recognition [12, 25, 47, 48, 50]. For the integration of various modalities, the most intuitive way is the fusion at the feature level. In feature-level fusion [16, 35, 41, 45, 53, 56], facial and vocal features are concatenated to construct a joint feature vector, and are then modeled by a single classifier for emotion recognition as shown in Fig. 4. For instance, Rosas *et al.* [53] used the SVMs with a linear kernel as the early fusion technique for binary classification. The experiments performed on the Spanish Multimodal Opinion database show that the integration of audio and visual features can improve significantly over the use of one modality at a time. To recognize continuously valued affective dimensions, Schuller *et al.* [16] concatenated the audio and video features into a single feature vector and used the support vector regression (SVR) as the baseline in the AVEC 2012 challenge. Eyben *et al.* [45] investigated an audiovisual fusion approach to classification of vocal outbursts (non-linguistic vocalizations) in noisy conditions. The visual features are complementary information, which is useful when the audio channel is noisy. Then the audio and visual modalities are fused at the feature level and classification is performed using Long Short-Term Memory Recurrent Neural Networks (LSTM-RNNs). Sayedelahl *et al.* [41] also proposed a fusion approach to enhance the recognition performance of continuous-valued emotion in spontaneous conversations. First, the audio features extracted from the whole utterance was concatenated with the visual features extracted from each of the individual visual frames with respect to the sentence, and the Supervised Principal Component Analysis (SPCA) was used to reduce the dimensions of the prosodic, spectral, and the facial features. Then a frame-level regression model was developed to estimate the continuous values of the three emotional dimensions (valence, activation, and dominance). Finally, a simple decision aggregation rule was used by averaging the resulting estimates of all image frames for final emotion recognition. Although fusion at feature level using simple concatenation of the audiovisual features has been successfully used in several applications, high-dimensional feature set may easily suffer from the problem of data sparseness, and does not take into account the interactions between features. Hence, the advantages of combining audio and visual cues at the feature level will be limited.

To eliminate the disadvantage of feature-level fusion strategy, a vast majority of research on data fusion strategies was explored toward the decision-level fusion. In decision-level fusion [15, 34, 36, 48, 49], multiple signals can be modeled by the corresponding classifier first, and then the recognition results from each classifier are fused in the end, as shown in Fig. 5. The fusion-based method at the decision level, without increasing the dimensionality, can combine various modalities by exploring the contributions of different emotional expressions. In the AVEC 2011 challenge [15], Schuller *et al.* first obtained predictions of the audio and video classifiers separately, then fused the two modalities by concatenating the two posterior probabilities and used a linear SVM as the audiovisual challenge baseline. Ramirez *et al.* [48] presented the late fusion using Latent-Dynamic Conditional Random Field (LDCRF) as a model to fuse the outputs of uni-modal classifiers. Error Weighted Classifier (EWC) combination [34, 36] is another well-known example. For EWC, Metallinou *et al.* [36] applied a Bayesian framework to combine empirical evidences with prior beliefs to fuse multiple cues. For voice modality, an HMM was trained for each emotional phonetic category. For face modality, the upper face is modeled by GMMs trained for each emotion with no viseme information, and the lower face is modeled by HMMs trained for each emotional viseme. Then the weighted sum of the individual decisions was combined to effectively combine various

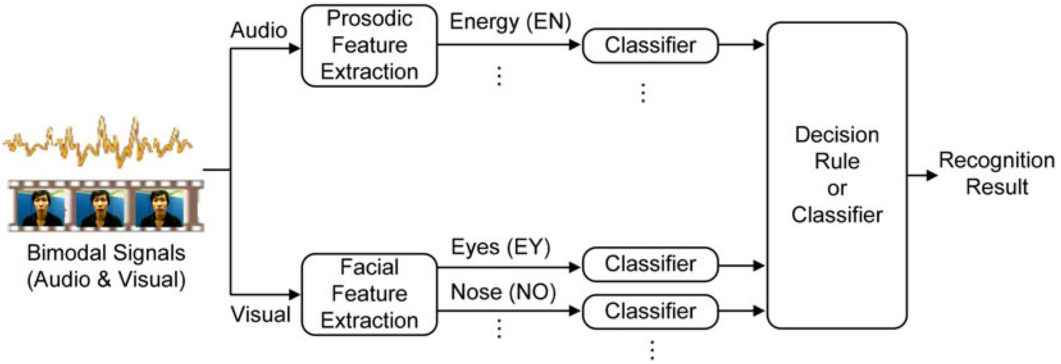

Audio
Prosodic Feature Extraction
Energy (EN)
Classifier
Bimodal Signals (Audio & Visual)
Decision Rule or Classifier
Recognition Result
Facial Feature Extraction
Eyes (EY)
Classifier
Nose (NO)
Classifier
Visual

Fig. 5. Illustration of decision-level fusion strategy for audiovisual emotion recognition.

modalities by exploring their contributions estimated from the confusion matrices of each classifier for final decision. However, as facial and vocal features have been confirmed complementary to each other in emotional expression [1], the assumption of conditional independence among multiple modalities at the decision level is inappropriate. The correlations between audio and visual modalities should be considered.

To deal with this problem, a model-level fusion strategy [31, 32, 50, 51, 75–77] was proposed to emphasize the information of correlation among multiple modalities, and explore the temporal relationship between audio and visual signal streams (as shown in Fig. 6). There are several distinctive examples such as Coupled HMM (C-HMM) [77, 98], Tripled HMM (T-HMM) [75], Multistream Fused HMM (MFHMM) [76], and Semi-Coupled HMM (SC-HMM) [51]. In C-HMM, which is a traditional example, two component HMMs are linked through cross-time and cross-chain conditional probabilities. This structure models the asynchrony of audio and visual modalities and preserves their natural correlations over time. Although cross-time and cross-chain causal modeling in C-HMM may better capture the inter-process influences between audio and visual modalities in real-world scenarios, a complex model structure and rigorous parameter estimation method of the statistical dependencies in C-HMM may lead to the overfitting effect in sparse data conditions. Further, Lu *et al.* [77] designed an AdaBoost-CHMM strategy which boosts the performance of component C-HMM classifiers with the modified expectation maximization (EM) training algorithm to generate a strong ensemble classifier. Song *et al.* [75] extended C-HMM to T-HMM to collect three HMMs for two visual input sequences and one audio sequence. Similarly, Jiang *et al.* [32] proposed a Triple stream audio visual Asynchronous Dynamic Bayesian Network (T_AsyDBN) to combine the MFCC features, local prosodic features and visual emotion features in a reasonable manner. Different from C-HMM and T-HMM, Zeng *et al.* [76] proposed the Multistream Fused HMM (MFHMM) which constructed a new structure linking the multiple component HMMs to detect 11 affective states. The MFHMM allows the building of an optimal connection among multiple streams according to the maximum entropy principle and the maximum mutual information criterion. To obtain a better statistical dependency among various modalities and diminish the overfitting effect, Lin *et al.* [51] proposed a novel connection criterion of model structures, which is a simplified state-based bi-modal alignment strategy in SC-HMM to align the temporal relation of the states between audio and visual streams.

On the one hand, a more sophisticated fusion strategy called hybrid approach was recently proposed to integrate different fusion approaches to obtain a better emotion recognition result. The Error Weighted SC-HMM (EWSC-HMM) [47], as an example of the hybrid approach, consists of model-level and decision-level fusion strategies and concurrently combines both advantages. First, the state-based bimodal alignment strategy in SC-HMM (model-level fusion) was proposed to align the temporal relation between audio and visual streams. Then the Bayesian classifier weighting scheme (decision-level fusion) was adopted to explore the contributions of the SC-HMM-based classifiers for different audio-visual feature pairs to obtain the optimal emotion recognition result.

D) A few related issues

Another important issue in audiovisual data fusion is related to the problem of asynchrony between audio and visual signals. From the speech production point of view, it has been proven that visual signal activity usually precedes the audio signal by as much as 120 ms [99, 100]. When people speak something happily, smile expression normally shows on faces earlier than the happy sound [97]. Hence, data fusion strategy will face the problem related to how to deal with asynchronous signals for audiovisual emotion recognition. For audiovisual data fusion, the current feature-level fusion methods dealt with the asynchrony problem based on a strict constraint on time synchrony between modalities or using the static features from each input utterance (i.e. ignoring the temporal information). Hence, with the assumption of strict time synchrony, feature-level fusion is unable to work well if the input features of the vocal and facial expressions differ in the temporal characteristics [12]. In addition, since decision-level fusion method focused on exploring how to effectively

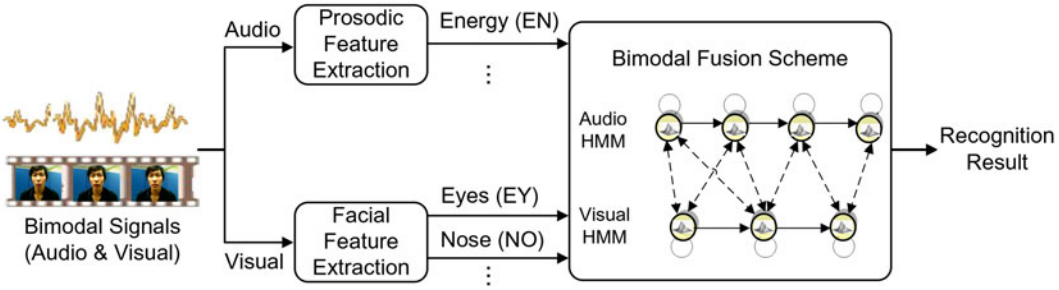


Fig. 6. Illustration of model-level fusion strategy for audiovisual emotion recognition.

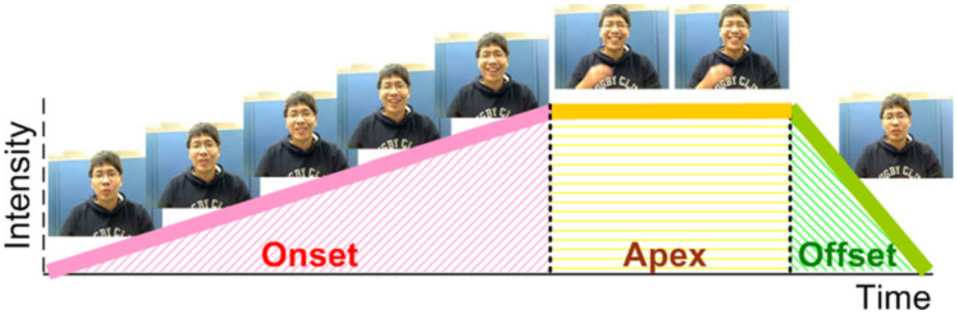


Fig. 7. An example of the temporal phases of happy facial expression from onset, over apex to offset phase.

combine the recognition outputs from individual audio and visual classifiers, which model audio and visual signal streams independently, synchronization issue can be ignored in decision-level fusion. On the one hand, the model-level fusion methods (e.g. C-HMM, T-HMM, SC-HMM, T_AsyDBN, etc. [32, 51, 75, 77]) were recently proposed and applied for audiovisual emotion recognition, trying to model asynchronous vocal and facial expressions, and preserving their natural correlation over time. Different from the dynamic programming algorithms (Viterbi and forward–backward analysis) used in conventional HMMs to handle temporal variations, the current model-level fusion methods [32, 51, 75, 77] were extended to deal with the synchronization problem by de-synchronizing the audio and visual streams and aligning the audiovisual signals at the state level. Thus the current model-level fusion methods such as C-HMM can achieve good performance for the audiovisual signals with large deviations in synchrony. In other distinctive examples, Song *et al.* [75] proposed a T-HMM-based emotion recognition system to model the correlations of three component HMMs, allowing unconstrained state asynchrony between these streams. Chen *et al.* [97] proposed an audiovisual DBN model with constrained asynchrony for emotion recognition, which allows asynchrony between the audio and visual emotional states within a constraint. Based on the mentioned studies, the most current model-level fusion methods tried to model the natural correlations between asynchronous vocal and facial expressions over time by exploring the relationships at "state" level.

Besides, toward naturalistic emotion recognition, several existing fusion strategies explored the evolution patterns of emotional expression in a conversational environment [35, 50]. These approaches considered the emotional substate or emotional state transitions within/between sentences in a conversation, which not only employed the correlation between audio and visual streams but also explored emotional sub-state or emotional state evolution patterns. Previous research has demonstrated that a complete emotional expression can be divided into three sequential temporal phases, onset (application), apex (release), and offset (relaxation), which consider the manner and intensity of an expression [101–104]. An example of the temporal phases of onset, apex, and offset of facial expression is shown in Fig. 7. In the onset phase of the example, the muscles contract and the appearance of the face changes as the facial action grows stronger. The apex phase represents that the facial action is at its peak and there are no more changes in facial appearance. In the offset phase, the muscles relax and the face returns to its neutral appearance. To this end, a bimodal HMM-based emotion recognition scheme, constructed in terms of emotional substates defined to represent temporal phases of onset, apex, and offset, was proposed to model the temporal course of an emotional expression for audio and visual signal streams. Wu *et al.* [50] proposed a Two-level Hierarchical alignment-based SC-HMM (2H-SC-HMM) fusion method to align the relationship within and between the temporal phases in the audio and visual HMM sequences at the state and model levels. Each HMM in the 2H-SC-HMM was used to characterize one emotional substate, instead of the entire emotional state. Figure 8 illustrates model- and state-level alignments between audio and visual HMM sequences in the happy emotional state. Furthermore, by integrating an emotional sub-state language model, which considers the temporal transition between emotional substates, the

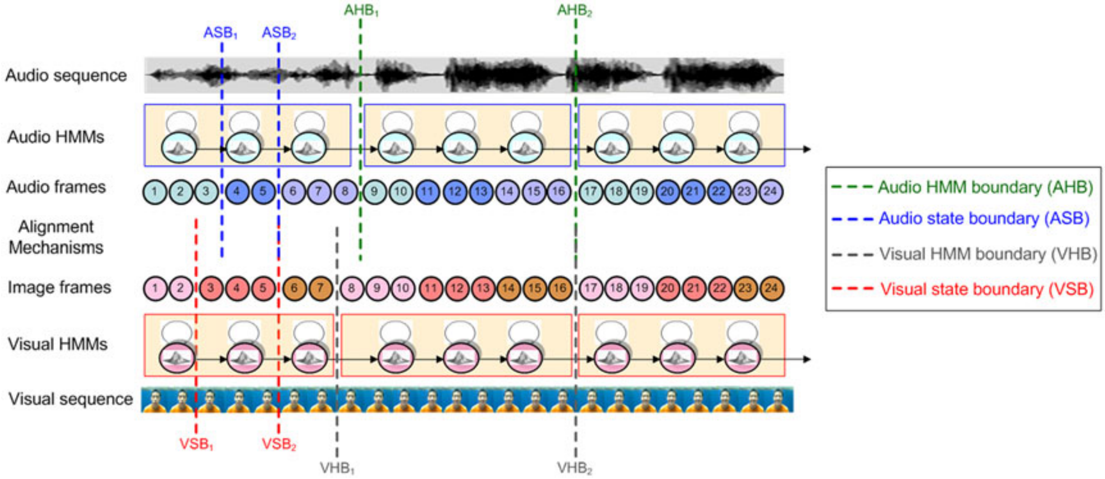


Fig. 8. An example illustrating model- and state-level alignments between audio and visual HMM sequences in the happy emotional state. The green and gray dotted lines represent the audio and visual HMM boundaries respectively and are used for model-level alignment estimation; the blue and red dotted lines represent the state boundaries under audio and visual HMMs respectively and are used for the state-level alignment estimation. The audio and image frames are represented by the numbered circles [50].

2H-SC-HMM can provide a constraint on allowable temporal structures to determine the final emotional state. In addition, Metallinou *et al.* [35, 105] explored toward the evolution patterns of emotional expression in a conversational environment that considers the emotional state transitions between utterances in a dialog. For example, anger to anger are more probable than anger to happiness. To model emotion evolution within an utterance and between utterances over the course of a dialog, the Bidirectional Long Short-Term Memory (BLSTM) networks was proposed for incorporating the past and future contextual information in audio-visual emotion recognition system. On the one hand, Mariooryad *et al.* [37] explored new directions to understand the emotional entrainment effect during dyadic spontaneous interactions. The relationship between acoustic features of the speaker and facial expressions of the interlocutor (i.e. cross-modality entrainment) was analyzed using mutual information framework. In IEMOCAP and SEMAINE databases, the results demonstrated the cross-modality and cross-speaker emotion recognition mechanism (i.e. recognizes the listener's emotions using facial features; recognizes the speaker's emotions using acoustic features) can improve the performance.

IV. CONCLUSION

This paper provides a survey on the latest research and challenges focusing on the theoretical background, databases, features, and data fusion strategies in audiovisual emotion recognition. First, the importance of integrating the facial and vocal cues is introduced. Second, we list the audiovisual emotion databases between 2006 and 2014 which were commonly used in audiovisual emotion recognition studies and emotion challenges from facial and vocal expressions. The content of the elicitation method and emotion categorization of the audiovisual emotion databases are also described. Third, the studies of data fusion strategies for facial–vocal expression-based emotion recognition in recent years are summarized, where the content of audio features, facial features, audiovisual bimodal fusion approach, and a few related issues are explained. Although a number of promising studies have been proposed and successfully applied to various applications, there are still some important issues, outlined in the following, needed to be addressed.

1. Unlike traditional emotion recognition performed on laboratory controlled data, EmotiW 2013 challenge provided a new direction to explore the performance of emotion recognition methods that work in real-world conditions.
2. A comprehensive and accessible database covering various social signals such as laughs [55], smiles, depression [17], etc. is desirable to help better understand affective behaviors.
3. For effective emotion recognition, more emotion-related information should be considered, such as textual (i.e. speech content) or body gesture information.
4. For features normalization, most studies assumed that the speaker ID was known, and the neutral example was often manually selected from the video sequences at the beginning. This assumption results in a limit to real-life applications and could be relaxed by automatically detecting the neutral segments [106] from the test data through a universal neutral model and a speaker identification system. The automatic feature normalization approach is a critical issue and should be considered in the future.
5. Compared with the unimodal methods, combining audio and visual cues can improve the performance of emotion recognition. Developing better data fusion approaches such as considering the various model properties, temporal expression, and asynchrony issue is desirable for multimodal emotion recognition to achieve better performance.

6. Exploring the expression styles from different users is an essential topic for effective emotion recognition, which is not only related to the expression intensity, but also related to the expression manner and significantly associated with personality trait.
7. Building a general emotion recognition system that performs equally well for every user could be insufficient for real applications. In contrast, it would be more desirable for personalized emotion recognition using personal computer/devices. Toward personalized emotion recognition, model adaptation based on a small-sized adaptation database should be considered in the future.
8. To increase the system's value in real-life applications, several existing methods tried to explore the issues on variations in spontaneous emotion expressions, including head pose variations, speaking-influenced facial expression, and partial facial occlusion in facial emotion recognition [107–109]. Investigations on these effects are essential for achieving robust emotion recognition.

ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Contract NSC102-2221-E-006-094-MY3 and the Headquarters of University Advancement at the National Cheng Kung University, which is sponsored by the Ministry of Education, Taiwan.

Supplementary Methods and Materials

The supplementary material for this article can be found at http://www.journals.cambridge.org/SIP

REFERENCES

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

[2] Cowie, R. *et al.*: Emotion recognition in human-computer interaction. IEEE Signal Process. Mag., 18 (2001), 33–80.

[3] Fragopanagos, N.; Taylor, J.G.: Emotion recognition in human-computer interaction. Neural Netw., 18 (2005), 389–405.

[4] Mehrabian, A.: Communication without words. Psychol. Today, 2 (1968), 53–56.

[5] Ambady, N.; Weisbuch, M.: Nonverbal behavior, in S.T. Fiske, D.T. Gilbert & G. Lindzey (Eds), John Wiley & Sons, Inc., Handbook of Social Psychology, 2010, 464–497.

[6] Rule, N.; Ambady, N.: First impressions of the face: predicting success. Social Person. Psychol. Compass, 4 (2010), 506–516.

[7] Devillers, L.; Vidrascu, L.: Real-life emotions detection with lexical and paralinguistic cues on human–human call center dialogs, in *Interspeech*, 2006, 801–804.

[8] Ayadi, M.E.; Kamel, M.S.; Karray, F.: Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognit., 44 (2011), 572–587.

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

[10] Sumathi, C.P.; Santhanam, T.; Mahadevi, M.: Automatic facial expression analysis a survey. Int. J. Comput. Sci. and Eng. Surv. (IJCSES), 3 (2013), 47–59.

[11] Pantic, M.; Bartlett, M.: Machine Analysis of Facial Expressions. Face Recognition. *I-Tech Education and Publishing*, Vienna, Austria, 2007, 377–416.

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

[13] Sebe, N.; Cohen, I.; Gevers, T.; Huang, T.S.: Emotion recognition based on joint visual and audio cues, in *Proc. 18th Int. Conf. Pattern Recognition*, 2006, 1136–1139.

[14] Busso, C. *et al.*: Analysis of emotion recognition using facial expression, speech and multimodal information, in *Proc. Sixth ACM Int'l Conf. Multimodal Interfaces*, 2004, 205–211.

[15] Schuller, B.; Valstar, M.; Eyben, F.; McKeown, G.; Cowie, R.; Pantic, M.: AVEC 2011 the first international audio/visual emotion challenge, in *Proc. First Int. Audio/Visual Emotion Challenge and Workshop (ACII)*, 2011, 415–424.

[16] Schuller, B.; Valstar, M.; Eyben, F.; Cowie, R.; Pantic, M.: AVEC 2012 – the continuous audio/visual emotion challenge, in *Proc. of Int. Audio/Visual Emotion Challenge and Workshop (AVEC)*, ACM ICMI, 2012.

[17] Valstar, M. *et al.*: AVEC 2013 – The Continuous Audio/Visual Emotion and Depression Recognition Challenge, *ACM Multimedia*, 2013.

[18] Valstar, M.*et al.*: AVEC 2014 – 3D dimensional affect and depression recognition challenge, in Proc. AVEC 2014, held in Conjunction with the 22nd ACM Int. Conf. Multimedia (MM 2014), 2014.

[19] Schuller, B.; Steidl, S.; Batliner, A.: The INTERSPEECH 2009 emotion challenge, in *Proc. Interspeech*, 2009, 312–315.

[20] Schuller, B. *et al.*: The INTERSPEECH 2010 paralinguistic challenge, in *Proc. INTERSPEECH*, 2010, 2794–2797.

[21] Schuller, B. *et al.*: The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism, in *Proc. Interspeech*, 2013, 148–152.

[22] Valstar, M.; Jiang, B.; Mehu, M.; Pantic, M.; Scherer, K.: The first facial expression recognition and analysis challenge, in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, 2011, 921–926.

[23] Dhall, A.; Goecke, R.; Joshi, J.; Wagner, M.; Gedeon, T.: Emotion recognition in the wild challenge 2013, in *ACM ICMI*, 2013.

[24] http://emotion-research.net/wiki/Databases

[25] D'Mello, S.; Kory, J.: Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies, in *Proc. ACM Int. Conf. Multimodal Interaction (ICMI)*, 2012, 31–38.

[26] Stuhlsatz, A.; Meyer, C.; Eyben, F.; Zielke, T.; Meier, G.; Schuller, B.: Deep neural networks for acoustic emotion recognition: raising the benchmarks, in *ICASSP*, 2011, 5688–5691.

[27] Gunes, H.; Schuller, B.: Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image Vis. Comput., 31 (2013), 120–136.

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

[29] Bänziger, T.; Pirker, H.; Scherer, K.: Gemep – Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions, in *Proc. of LREC Workshop on Corpora for Research on Emotion and Affect*, 2006, 15–19.

[30] Martin, O.; Kotsia, I.; Macq, B.; Pitas, I.: The eNTERFACE'05 audio-visual emotion database, in *Int. Conf. Data Engineering Workshops*, 2006.

[31] Paleari, M.; Benmokhtar, R.; Huet, B.: Evidence theory-based multimodal emotion recognition, in *Proc. 15th Int. Multimedia Modeling Conf. Advances in Multimedia Modeling*, 2009, 435–446.

[32] Jiang, D.; Cui, Y.; Zhang, X.; Fan, P.; Gonzalez, I.; Sahli, H.: Audio visual emotion recognition based on triple-stream dynamic Bayesian network models, in *Proc. Affective Computing and Intelligent Interaction*, 2011, 609–618.

[33] Busso, C. *et al.*: IEMOCAP: interactive emotional dyadic motion capture database. J. Lang. Resources Eval., 42 (2008), 335–359.

[34] Metallinou, A.; Lee, S.; Narayanan, S.: Audio-visual emotion recognition using Gaussian mixture models for face and voice in *Proc. Int. Symp. Multimedia*, 2008, 250–257.

[35] Metallinou, A.; Wollmer, M.; Katsamanis, A.; Eyben, F.; Schuller, B.; Narayanan, S.: Context-sensitive learning for enhanced audiovisual emotion classification. IEEE Trans. Affective Comput., 3 (2012), 184–198.

[36] Metallinou, A.; Lee, S.; Narayanan, S.: Decision level combination of multiple modalities for recognition and analysis of emotional expression, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, 2010, 2462–2465.

[37] Mariooryad, S.; Busso, C.: Exploring cross-modality affective reactions for audiovisual emotion recognition. IEEE Trans. Affective Comput., 4 (2013), 183–196.

[38] Wang, Y.; Guan, L.: Recognizing human emotional state from audiovisual signals. IEEE Transactions on Multimedia, 10 (2008), 936–946.

[39] Wang, Y.; Zhang, R.; Guan, L.; Venetsanopoulos, A.N.: Kernel fusion of audio and visual information for emotion recognition, in *Proc. 8th Int. Conf. Image Analysis and Recognition (ICIAR)*, 2011, 140–150.

[40] Grimm, M.; Kroschel, K.; Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database, in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2008, 865–868.

[41] Sayedelahl, A.; Araujo, P.; Kamel, M.S.: Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations, in *Int. Conf. Multimedia and Expo Workshops*, 2013, 1–6.

[42] Haq, S.; Jackson, P.J.B.: Speaker-dependent audio-visual emotion recognition, in *Proc. Int. Conf. Auditory-Visual Speech Processing*, 2009, 53–58.

[43] Haq, S.; Jackson, P.J.B.: Multimodal emotion recognition, in W. Wang (ed), Machine Audition: Principles, Algorithms and Systems, *IGI Global Press*, chapter 17 (2010), 398–423.

[44] Schuller, B.; Müller, R.; Hörnler, B.; Höthker, A.; Konosu, H.; Rigoll, G.: Audiovisual recognition of spontaneous interest within conversations, in *Proc. 9th Int. Conf. Multimodal Interfaces (ICMI), Special Session on Multimodal Analysis of Human Spontaneous Behaviour*, ACM SIGCHI, 2007, 30–37.

[45] Eyben, F.; Petridis, S.; Schuller, B.; Pantic, M.: Audiovisual vocal outburst classification in noisy acoustic conditions, in *ICASSP*, 2012, 5097–5100.

[46] McKeown, G.; Valstar, M.; Pantic, M.; Cowie, R.: The SEMAINE corpus of emotionally coloured character interactions, in *Proc. IEEE Int. Conf. Multimedia & Expo*, 2010, 1–6.

[47] Lin, J.C.; Wu, C.H.; Wei, W.L.: Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. IEEE Trans. Multimedia, 14 (2012), 142–156.

[48] Ramirez, G.A.; Baltrušaitis, T.; Morency, L.P.: Modeling latent discriminative dynamic of multi-dimensional affective signals, in *Proc. Affective Computing and Intelligent Interaction*, 2011, 396–406.

[49] Wöllmer, M.; Kaiser, M.; Eyben, F.; Schuller, B.; Rigoll, G.: LSTM-modeling of continuous emotions in an audiovisual affect recognition framework, in Image and Vision Computing (IMAVIS). Spec. Issue Affect Anal. Continuous Input, 31 (2013), 153–163.

[50] Wu, C.H.; Lin, J.C.; Wei, W.L.: Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course. IEEE Trans. Multimedia, 15 (2013), 1880–1895.

[51] Lin, J.C.; Wu, C.H.; Wei, W.L.: Semi-coupled hidden Markov model with state-based alignment strategy for audio-visual emotion recognition, in *Proc. Affective Computing and Intelligent Interaction (ACII)*, 2011, 185–194.

[52] Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T.: Collecting large, richly annotated facial-expression databases from movies. IEEE Multimedia, 19 (2012), 34–41.

[53] Rosas, V.P.; Mihalcea, R.; Morency, L.-P.: Multimodal sentiment analysis of Spanish online videos. IEEE Intell. Syst., 28 (2013), 38–45.

[54] Morency, L.-P.; Mihalcea, R.; Doshi, P.: Towards multimodal sentiment analysis: harvesting opinions from the web, in *Proc. 13th Int. Conf. Multimodal Interfaces (ICMI)*, 2011, 169–176.

[55] Petridis, S.; Martinez, B.; Pantic, M.: The MAHNOB laughter database. Image Vis. Comput., 31 (2013), 186–202.

[56] Rudovic, O.; Petridis, S.; Pantic, M.: Bimodal log-linear regression for fusion of audio and visual features, in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, 789–792.

[57] Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. 2nd Int. Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE), in *Proc. IEEE Face & Gestures*, 2013, 1–8.

[58] Bänziger, T.; Scherer, K.: Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) Corpus. *Oxford University Press*, Oxford, 2010, 271–294.

[59] Bänziger, T.; Mortillaro, M.; Scherer, K.: Introducing the Geneva multimodal expression corpus for experimental research on emotion perception. Emotion, 12 (2012), 1161–1179.

[60] http://www.enterface.net/enterface05/main.php?frame=emotion

[61] Douglas-Cowie, E.; Cowie, R.; Cox, C.; Amier, N.; Heylen, D.: The sensitive artificial listener: an induction technique for generating emotionally coloured conversation, in *LREC Workshop on Corpora for Research on Emotion and Affect*, 2008, 1–4.

[62] Weizenbaum, J.: ELIZA – a computer program for the study of natural language communication between man and machine. Commun. ACM, 9 (1966), 36–45.

[63] Mckeown, G.; Valstar, M.F.; Cowie, R.; Pantic, M.; Schroe, M.: The SEMAINE database: annotated multimodal records of emotionally coloured conversations between a person and a limited agent. IEEE Trans. Affective Comput., 3 (2012), 5–17.

[64] http://www.semaine-db.eu/

[65] Ekman, P.; Friesen, W.V.: Picture of Facial Affect. *Consulting Psychologist Press*, Palo Alto, 1976.

[66] Ekman, P.: Facial expression and emotion, Am. Psychol., 48 (1993), 384–392.

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

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

[69] Thayer, R.E.: The Biopsychology of Mood and Arousal. *Oxford University Press*, New York, 1989.

[70] Yang, Y.-H.; Lin, Y.-C.; Su, Y.-F.; Chen, H.-H.: A regression approach to music emotion recognition. IEEE Trans. Audio, Speech Lang. Process., 16 (2008), 448–457.

[71] Zeng, Z.; Zhang, Z.; Pianfetti, B.; Tu, J.; Huang, T.S.; Audio-visual affect recognition in activation-evaluation space, in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2005, 828–831.

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

[73] Schuller, B. *et al.*: Being bored? Recognising natural interest by extensive audiovisual integration for real-life application. Image Vis. Comput. J., 27 (2009), 1760–1774.

[74] http://avec2013-db.sspnet.eu

[75] Song, M.; You, M.; Li, N.; Chen, C.: A robust multimodal approach for emotion recognition. Neurocomputing, 71 (2008), 1913–1920.

[76] Zeng, Z.; Tu, J.; Pianfetti, B.M.; Huang, T.S.: Audio-visual affective expression recognition through multistream fused HMM. IEEE Trans. Multimedia, 10 (2008), 570–577.

[77] Lu, K.; Jia, Y.: Audio-visual emotion recognition with boosted coupled HMM, in *Int'l Conf. Pattern Recognition (ICPR)*, 2012, 1148–1151.

[78] Wu, C.H.; Yeh, J.F.; Chuang, Z.J.: Emotion Perception and Recognition from Speech, Affective Information Processing, *Springer*, New York, 2009, 93–110.

[79] Wu, C.H.; Liang, W.B.: Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. IEEE Trans. Affective Comput., 2 (2011), 1–12.

[80] Morrison, D.; Wang, R.; De Silva, L.C.: Ensemble methods for spoken emotion recognition in call-centres. Speech Commun., 49 (2007), 98–112.

[81] Murray, I.R.; Arnott, J.L.: Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. J. Acoust. Soc. Am., 93 (1993), 1097–1108.

[82] Scherer, K.R.: Vocal communication of emotion: a review of research paradigms. Speech Commun., 40 (2003), 227–256.

[83] Luengo, I.; Navas, E.; Hernáez, I.; Sánchez, J.: Automatic emotion recognition using prosodic parameters, in *Proc. INTERSPEECH*, 2005, 493–496.

[84] Kooladugi, S.G.; Kumar, N.; Rao, K.S.: Speech emotion recognition using segmental level prosodic analysis, in *Int. Conf. Devices and Communications*, 2011, 1–5.

[85] Kwon, O.W.; Chan, K.; Hao, J.; Lee, T.W.: Emotion recognition by speech signals, in *Proc. 8th European Conf. Speech Comm. and Technology*, 2003.

[86] Freedman, D.A.: Statistical Models: Theory and Practice. *Cambridge University Press*, Cambridge, 2005.

[87] Boersma, P.; Weenink, D.: Praat: doing phonetics by computer, 2007. [Online]. Available: http://www.praat.org/.

[88] Eyben, F.; Wöllmer, M.; Schuller, B.: OpenSMILE – the Munich versatile and fast open-source audio feature extractor, In *Proc. 9th ACM Int. Conf. Multimedia*, MM, 2010, 1459–1462.

[89] Eyben, F.; Wöllmer, M.; Schuller, B.: OpenEAR introducing the Munich opensource emotion and affect recognition toolkit, in *Proc. Affective Computing and Intelligent Interaction (ACII)*, 2009, 576–581.

[90] Huang, Y.; Zhang, G.; Li, X.; Da, F.: Improved emotion recognition with novel global utterance-level features. Int. J. Appl. Math. Inf. Sci., 5 (2011), 147–153.

[91] Schuller, B.; Steidl, S.; Batliner, A.; Schiel, F.; Krajewski, J.: The INTERSPEECH 2011speaker state challenge, in *Proc. Interspeech*, 2011, 3201–3204.

[92] Patras, I.; Pantic, M.: Particle filtering with factorized likelihoods for tracking facial features, in *Proc. FG*, 2004, 97–104.

[93] Cootes, T.F.; Edwards, G.J.; Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell., 23 (2001), 681–685.

[94] Shan, C.; Gong, S.; Mcowan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput., 27 (2009), 803–816.

[95] Ahonen, T.; Hadid, A.; Pietikäinen, M.: Face description with local binary patterns: application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell., 28 (2006), 2037–2041.

[96] OKAO: software http://www.omron.com/r_d/coretech/vision/okao.html

[97] Chen, D.; Jiang, D.; Ravyse, I.; Sahli, H.: Audio-visual emotion recognition based on a DBN model with constrained asynchrony, in *Fifth Int. Conf. Image and Graphics*, 2009, 912–916.

[98] Nicolaou, M.; Gunes, H.; Pantic, M.: Audio-visual classification and fusion of spontaneous affective data in likelihood space, in *Int. Conf. Pattern Recognition (ICPR)*, 2010, 3695–3699.

[99] Grant, K.W.; Greenberg, S.: Speech intelligibility derived from asynchronous processing of auditory-visual information, in *Proc. Workshop on Audio-Visual Speech Processing (AVSP)*, 2001, 132–137.

[100] Xie, L.; Liu, Z.Q.: A coupled HMM approach to video-realistic speech animation. Pattern Recognit., 40 (2007), 2325–2340.

[101] Valstar, M.F.; Pantic, M.: Fully automatic recognition of the temporal phases of facial actions. IEEE Trans. Syst. Man Cybern. B, 42 (2012), 28–43.

[102] Koelstra, S.; Pantic, M.; Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. IEEE Trans. Pattern Anal. Mach. Intell., 32 (2010), 1940–1954.

[103] Jiang, B.; Valstar, M.; Martinez, B.; Pantic, M.: A dynamic appearance descriptor approach to facial actions temporal modeling. IEEE Trans. Syst. Man Cybern. B, 44 (2014), 161–174.

[104] Lin, J.C.; Wu, C.H.; Wei, W.L.: Emotion recognition of conversational affective speech using temporal course modeling, in *Proc. Interspeech*, 2013, 1336–1340.

[105] Wöllmer, M.; Metallinou, A.; Eyben, F.; Schuller, B.; Narayanan, S.S.: Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling, in *INTERSPEECH*, 2010, 2362–2365.

[106] Busso, C.; Metallinou, A.; Narayanan, S.: Iterative feature normalization for emotional speech detection, in *Proc. Int. Conf. Acoust., Speech, and Signal Processing*, 2011, 5692–5695.

[107] Rudovic, O.; Pantic, M.; Patras, I.: Coupled Gaussian processes for pose-invariant facial expression recognition. IEEE Trans. Pattern Anal. Mach. Intell., 35 (2013), 1357–1369.

[108] Wu, C.H.; Wei, W.L.; Lin, J.C.; Lee, W.Y.: Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion. IEEE Trans. Multimedia, 15 (2013), 1732–1744.

[109] Lin, J.C.; Wu, C.H.; Wei, W.L.: Facial action unit prediction under partial occlusion based on error weighted cross-correlation model, in *Int. Conf. Acoustics, Speech, and Signal Processing*, 2013, 3482–3486.

Chung-Hsien Wu received the Ph.D. degree in Electrical Engineering from National Cheng Kung University, Taiwan, in 1991. Since 1991, he has been with the Department of Computer Science and Information Engineering, National Cheng Kung University. He became the professor and distinguished professor in 1997 and 2004, respectively. He received the Outstanding Research Award of National Science Council in 2010 and the Distinguished Electrical Engineering Professor Award of the Chinese Institute of Electrical Engineering in 2011, Taiwan. He is currently the associate editor of IEEE Trans. Audio, Speech and Language Processing, IEEE Trans. Affective Computing, and ACM Trans. Asian Language Information Processing. Dr. Wu serves as the Asia Pacific Signal and Information Processing Association (APSIPA) Distinguished Lecturer and Speech, Language and Audio (SLA) Technical Committee Chair in 2013–2014. His research interests include multimodal emotion recognition, speech recognition/synthesis, and spoken language processing.

Jen-Chun Lin received the Ph.D. degree in Computer Science and Information Engineering from National Cheng Kung University, Tainan, Taiwan, Republic of China, in 2014. Currently, he is a postdoctoral fellow of the Institute of Information Science, Academia Sinica, Taiwan. He is a student member of IEEE and also a student member of HUMAINE Association (emotion-research.net). His research interests include multimedia signal processing, pattern analysis and recognition, and affective computing.

Wen-Li Wei received the B.S. and M.S. degrees in Information Engineering from I-Shou University, Kaohsiung, Taiwan, Republic of China, in 2006 and 2008, respectively. She is currently working toward the Ph.D. degree in Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, Republic of China She is a student member of IEEE. Her research interests include multimedia signal processing, pattern analysis and recognition, and affective computing.