Emotion Recognition from Multi-Modal Information

Chung-Hsien Wu, Jen-Chun Lin, Wen-Li Wei and Kuan-Chun Cheng
Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
E-mail: chunghsienwu@gmail.com, jenchunlin@gmail.com, lilijinjin@gmail.com, davidcheng817@gmail.com

***Abstract*—Emotion recognition is the ability to detect what people are feeling from moment to moment and to understand the connection between their feelings and verbal/non-verbal expressions. When you are aware of your emotions, you can think clearly and creatively, manage stress and challenges, communicate well with others, and display trust, empathy, and confidence. In today's world, human-computer interaction (HCI) interface undoubtedly plays an important role in our daily life. Toward harmonious HCI interface, automated analysis of human emotion has attracted increasing attention from the researchers in multidisciplinary research fields. In this paper, we presents a survey on theoretical and practical work offering new and broad views of the latest research in emotion recognition from multi-modal information including facial and vocal expressions. A variety of theoretical background and applications ranging from salient emotional features, emotional-cognitive models, to multi-modal data fusion strategies is surveyed for emotion recognition on these modalities. Conclusions outline some of the existing emotion recognition challenges.**

I. INTRODUCTION

Emotions are important in human intelligence, rational decision making, social interaction, perception, learning, memory and more [1]. Intact perception and experience of emotion is vital for communication in the social environment. Accordingly, understanding emotions is indispensable for the day-to-day functioning 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 technology provides an opportunity to promote harmonious interactions or communication between computers and humans [2]–[4]. Perception and production of emotional expression are central to verbal/non-verbal communication. Hence, constructing an emotion recognition system from multi-modal information is desirable.

Psychologists have various opinions about the importance of different cues in human affect judgment such as from facial expression, vocal expression or linguistic message. Among these analyses, facial expression is acknowledged as one of the most direct channels to transmit human emotions on non-verbal communication. Two main issues in the current research on automatic analysis of facial expressions consider facial affect (e.g., emotion) recognition and facial muscle action recognition [2], [5]–[7]. The former is to infer what psychological meaning of facial expression are displayed, such as happy or angry emotional state, and the latter is to describe the "surface" of the facial behavior, such as change of facial shape or appearance [8]. Most facial affect recognition issue attempts to recognize a small set of prototypical emotional facial expressions such as the six prototypical emotions: anger, disgust, fear, happiness, sadness and surprise proposed by Ekman [9], [10]. Even though the automatic facial affect recognition is well studied, prototypical emotions cover only a subset of the total range of possible facial displays and categorization of facial expressions. For example, the boredom and interest cannot seem to fit well in any of the prototypical emotions. To understand the change of subtle facial behaviors and human emotions, automatic recognition of atomic facial signals is needed. Accordingly, Facial Action Coding System (FACS) [11] was proposed to classify the atomic facial signals into Action Units (AUs) through analysis of facial muscle contractions. Of 44 AUs defined in the original FACS version [11] and revised in [12], 32 AUs were redefined [5], [8], [13] to represent the smallest visually discernible facial movements. Using FACS, human coders can manually code nearly any anatomically possible facial expression, and the AUs can be used as the basis for any higher order decision making process including the recognition of prototypical emotions, cognitive states, social signals, and more [2]. Thus, FACS motivated the studies on automatic human spontaneous facial behavior recognition.

Speech is another one of the most important and natural channels to transmit human affective states especially on verbal communication. Affective cues can convey the most important aspect of what is said, such as whether the speaker liked something or not [1]. Affective information in speech can be transmitted through explicit (linguistic) and implicit (paralinguistic) messages during communication [2]. The former can be understood and extracted from affective words, phrases, sentences, semantic contents, and more, and the later may be explored from prosodic and acoustic information of speech. Similar to facial expression analysis, most of the speech emotion recognition studies are focused on recognizing the six prototypical emotional states [2], [3], [14]. In addition, some of non-prototypical emotions were also explored such as frustration, puzzle, boredom, and bore based on the linguistic information or nonlinguistic vocalizations [15]–[17]. 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

Fig. 1. Examples of the AAM alignment results including eyebrows, eyes, nose, mouth, and facial contours.

Fig. 2. An example of the temporal phases of happy facial expression: (i) the onset phase, where the muscles are contracting and the appearance of the face changes as the facial action grows stronger, (ii) the apex phase, where the facial action is at its peak and there are no more changes in facial appearance, and (iii) the offset phase, where the muscles are relaxing and the face returns to its neutral appearance [8].

human emotion where an emotional state is characterized in numerous latent dimensions [18], [19]. According to the analysis from psychologists, a two-dimensional model (i.e., Arousal and Valence (A-V) represent active/passive and negative/positive affective responses, respectively) is deemed sufficient for capturing most affective variability. In addition, a promising area of research is the combination of facial and speech information to improve emotion recognition performance in both domains [1]. Hence, automatic, dimensional, and continuous emotion recognition from facial expression, speech, or both has increasingly attracted the interest of affective computing from the researchers in recent years [20]–[24].

In this paper, we first present a review of recent advances in the research on audio, visual, and audio-visual emotion recognition. The survey focuses on the fields of facial and vocal expression analysis ranging from exploring salient emotional features, emotional-cognitive models, to multi-modal data fusion strategies targeting the pattern recognition researchers who do not necessarily have a deep background in such areas.

The rest of the paper is organized as follows: Section II briefly outlines the state-of-the-art facial expression-based emotion/muscle action recognition studies. Section III reviews the state-of-the-art vocal expression-based emotion recognition studies. Section IV provides the reviews of the state-of-the-art facial-vocal expression-based emotion recognition studies. Section V offers the conclusion.

II. Facial Expression-Based Emotion/Muscle Action Recognition

Previous facial expression-based emotion/muscle action recognition studies generally employed the typical pattern recognition approaches, and were based on the 2D spatiotemporal facial features [2]. The commonly used facial feature types can be divided into two major categories including geometric and appearance features. The former represents the shape or location of facial components (i.e., eyebrows, eyes, mouth, etc.), and the later depicts the facial texture such as wrinkles, bulges, and furrows. The features of shape and location are estimated based on the results of facial component alignments through the classical approach "Active Appearance Model (AAM)" [25] as shown in Fig. 1. The facial appearance can be extracted from the methods, such as Gabor wavelets, spatial ratio face template, Haar feature, etc [26]–[28]. Given the extracted facial features, emotion and muscle action are then recognized by various pattern recognition methods [2], [29], such as Support Vector Machines (SVM), rule-based approach, AdaBoost classifiers, Sparse Representation (SR) classifiers, etc. In contrast to previous studies, the existing emotion and muscle action recognition approaches explore the structural and temporal characteristic for facial expression development [8], [30]–[32]. For structural and temporal characteristic modeling, Tong et al. employed a Dynamic Bayesian Network (DBN) to systematically model the spatiotemporal relationships among AUs which provide a coherent and unified hierarchical probabilistic framework to obtain AU measures, and achieved a significant performance improvement [30], [31]. Li et al. extended the Tong's approach, which creates a new hierarchical structure to model the facial interactions among facial feature point layer, AU layer, and expression (emotion) layer through DBN to improve the tracking and recognition performances of three layers simultaneously [32]. Valstar and Pantic proposed a combination of GentleBoost, SVM, and Hidden Markov Model (HMM) to encode AUs and their temporal activation which model the temporal characteristics of AUs for neutral, onset, apex, and offset temporal phases during expressions [8]. An example of the temporal phases of onset, apex, and offset of facial expression is shown in Fig. 2. In order to increase the system's value in real life applications, a part of the existing studies explore the issues on the effects of head pose variations, speaking-influenced facial expression, and partial facial occlusion on facial expression recognition studies [33]–[35]. Notable examples for mentioned effect are described as follows. For managing the effect of head pose variations, Rudovic et al. [33] proposed the Coupled Scaled Gaussian Process Regression (CSGPR) model for head-pose normalization, which first learns the mappings between the facial points in each pair of non-frontal and frontal pose, and then applies their coupling to capture the dependency between them. To solve the effect of speaking-influenced facial expression, Wu et al. [34] proposed an eigenface conversion-based approach to remove speaking effect on facial expressions. In the proposed approach, a context-dependent linear conversion function modeled by a statistical Gaussian Mixture Model (GMM) is constructed with parallel data from speaking and non-speaking facial expressions with emotions. The visual temporal context of the Articulatory Attribute (AA) classes is further considered for categorizing the conversion function through decision tree for precise modeling purpose. In terms of managing the effect of partial facial occlusion, Lin

TABLE I
FACIAL EXPRESSION-BASED EMOTION/MUSCLE ACTION RECOGNITION

References	Feature	Approaches	Class	Evaluated Database	Accuracy
Tong et al. [31]	• Gabor wavelets	• Adaboost • DBN	• Posed: 14 AUs • Spontaneous: 12 AUs	• Posed: CK [36] and ISL databases • Spontaneous: MAD, Belfast [37], and youtube	• 88.3 correct-positive rate (CK) • decrease positive-error rate to 24.3 (Spontaneous)
Li et al. [32]	• Gabor filter • feature points	• Adaboost • DBN	• 15 AUs • Six basic emotions	• CK+ database [38] - 309 sequences - 90 subjects • MMI database	• Recognition rate of AUs: 94.05 (CK+) • Recognition rate of emotions: 87.43 (CK+) • Recognition rate of AUs: 82.4 (MMI)
Valstar et al. [8]	• Tracked feature points	• GentleSVM • HMM	• 22 AUs • four temporal phases • six basic emotions	• CK • MMI [39] • DS118 databases	• Classification rate of AUs: - 91.7 (CK) - 95.3 (MMI) • F1-measure for neutral, onset, apex, and offset phases: 80.7, 53.7, 65.6, and 45.9 (MMI) • Emotion CR: 69.59 (CK)
Rudovic et al. [33]	• 39 landmarks	• CSGPR • SVM	• Seven emotions with head pose variations	• Posed: BU-3DFE [40], Multi-PIE [41], MPFE • Spontaneous: SEMAINE [42]	• Recognition rate for balanced data sets: - 76.5 (BU-3DFE) - 94.8 (Multi-PIE) - 79.1 (MPFE)
Wu et al. [34]	• Aligned facial feature points	• Eigenface Conversion • template matching • SVR	• Four emotion quadrant • Arousal-valence values	• Posed: extended MHMC audio-visual database • 6 actors • 2,160 sentences	• Recognition rate: 88.33 (four quadrants) • MAE: - Arousal 0.813 - Valence 0.671
Lin et al. [35]	• FDPs	• GMM-based EWCCM	• 5 AU and AU combination	• CK database • 90 subjects • 176 images with partial facial occlusion	• Prediction rate: 80.68

SVR: Support Vector Regression, MAE: Mean Absolute Error, FDPs: Facial Deformation Parameters.

et al. [35] proposed an Error Weighted Cross-Correlation Model (EWCCM) considering the cross-correlation among paired facial features and exploring their contributions to predict the AU under partial facial occlusion for emotion recognition.

The information related to the features, classifiers, and performances of the mentioned the-state-of-the-art approaches is illustrated in Table I.

III. VOCAL EXPRESSION-BASED EMOTION RECOGNITION

An important issue for emotion recognition from speech is the selection of relevant features. Several popular features such as prosodic and spectral features of emotional speech signals have been discussed over the years. 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 [2], [43]–[46]. Several studies [2], [47] have further noted that pitch- and energy-related features are useful to determine emotion in speech. In addition, Morrison et al. [46] further summarized the correlations between prosodic features and emotions. These findings conclude that prosody-related features are highly beneficial in emotion recognition. Besides, the spectral and voice quality features such as Mel-Frequency Cepstrum Coefficients (MFCC), Linear Predictor Coefficients (LPC), tense, and breathy were also frequently employed and discussed [3]. For speech emotion recognition, the speech features can be divided into two major categories including local and global features according to the model properties. The local features represent the speech features extracted based on the unit of "speech frame". On the other hand, the global features are calculated as the statistics of all speech features extracted from an "utterance" [3]. Based on the extracted speech features (i.e., local or global features), traditional pattern recognition methods such as dynamic modeling approach, HMM, or static modeling approaches, GMM, SVM, etc. have been used in almost all speech emotion recognition systems to decide the underlying emotion of the speech utterance. The dynamic modeling approach is applied to capture the temporal characteristics of affective speech and the detailed feature fluctuations, which can be trained reliably using a large number of local feature vectors. Although the dynamic modeling approaches are useful in considering the temporal information, the performance may be degraded on the effects from language, speaker, speech content, and especially from the complex temporal course of emotional expression of real conversational environment [24], [48]. In terms of static modeling approaches, a statistical global feature extraction method is employed to avoid the over-modeling problem, and is relatively insensitive to the effects of language, speaker, and speech content within the utterance. However, the change of emotion in time is short-lived and intense that ignored the temporal characteristics may also limit the recognition performance. Generally, these two modeling approaches have comparable performance of speech emotion recognition, yet the results are still unsatisfactory.

Due to the limitation mentioned above, recent studies began to investigate some of more appropriate units in addition to utterance and speech frame units. There are two strategies in unit selection. The first described here is defined

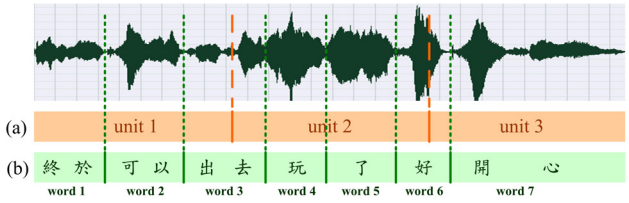

Fig. 3. Examples of (a) inflexible unit: utterance is subdivided into three parts of equal-length units and (b) flexible unit: word-based units.

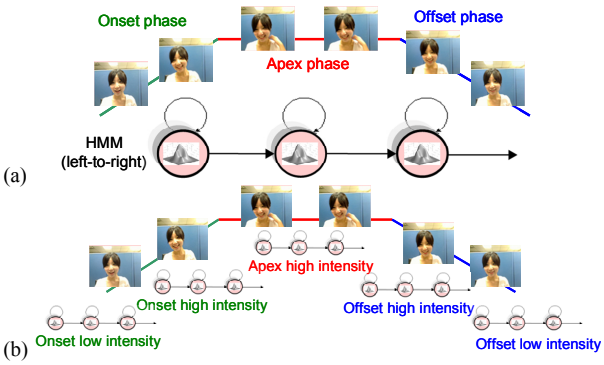

Fig. 4. (a) A single HMM is applied to model the entire temporal course of emotional expression (b) an HMM is used to model one temporal phase (sub-emotion state) of emotion expression.

as "inflexible units" [49], [50], such as fixed-length time slices or proportions of longer units (for example, utterance is subdivided into three parts of equal-length units) as shown in Fig. 3 (a). On the contrary, the second strategy is defined as "flexible units" with varying lengths such as syllables, words, phrases, or emotional salient segments [45], [50], [51], which are linguistically or meaningfully defined as shown in Fig. 3 (b). Moreover, most of the studies have shown that different units are providing different aspects on the affective speech and complementary to each other for recognition. In addition, some of the existing approaches explored the approaches for temporal evolution/course modeling. For example, Ntalampiras et al. [52] investigated three methods including short-term statistics, spectral moments, and autoregressive models for integrating subsequent feature values as well as modeling their evolution in time. Lin et al. [48] proposed a temporal course modeling approach and a sub-emotion language model, considering the temporal phases (i.e., onset, apex, and offset with high or low intensity) of an emotional expression, which tries to solve the problem of complex temporal course of emotional expression in a conversational environment. Instead of directly applied a single left-to-right topology of the HMM to model the entire temporal course of emotional expression (entire temporal phases of onset, apex, and offset) as shown in Fig. 4 (a), Lin used an HMM to characterize one sub-emotional state (temporal phase) as shown in Fig. 4 (b) and then applied the sub-emotion language model to provide a constraint on allowable temporal structures to obtain an optimal recognition result of emotional state of an utterance in a conversational environment. The formula of the temporal course modeling is described in the following:

$$\hat{E} = \arg\max_{E} P(E \mid O) = \arg\max_{E} P(O \mid E)P(E)$$
$$= \arg\max_{E} P(O \mid e_1, e_2, ..., e_M) \prod_{k=2}^{M} P(e_k \mid e_{k-1}) \quad (1)$$

where O represents the observation sequence $O = o_1^T = o_1, o_2, ..., o_T$, $\hat{E}$ represents the emotion recognition result with temporal phase sequence $e_1, e_2, ..., e_M$, and the probability of temporal phase transition is modeled using a bigram language model (so called sub-emotion language model in Lin's work) $P(E) = P(e_1, e_2, ..., e_M) = \prod_{k=2}^{M} P(e_k \mid e_{k-1})$.

The detailed information related to the utilized features, classifier, and performance of the inflexible- and flexible-units approaches as well as temporal evolution/course modeling approaches are summarized in Table II.

IV. Facial-Vocal Expression-Based Emotion Recognition

Humans have accessed to both visual and auditory channels in natural unmediated communication. Consequently, it is no surprise that the facial and vocal expressions are complementary to recognize human emotions, given that arousal is more usefully distinguished in speech, and valence is more easily distinguished in facial expressions [1]. To improve emotion recognition performance, a promising research area is to explore the data fusion strategy to effectively combine the facial and vocal cues. Accordingly, many data fusion strategies have been developed for facial-vocal expression-based emotion recognition in recent years. The fusion operations reported can be classified into four major categories: feature-level fusion, decision-level fusion, model-level fusion, and more recently, hybrid approach, for audio-visual emotion recognition [2], [23], [24].

For the integration of various modalities, the most intuitive way is to fuse them at the feature level. In feature-level fusion [55]–[57], facial and vocal features are concatenated to construct a joint feature vector, and are then modeled by a single classifier for emotion recognition. Although fusion at the feature level may obtain the advantage of combining visual and audio cues, high dimensional feature set may easily suffer from the problem of data sparseness, and stress the computational resources. To solve the disadvantage of feature-level fusion strategy, a vast majority of investigations on data fusion strategies are explored toward the decision-level fusion. In decision-level fusion [58]–[60], multiple signals can be modeled by the corresponding classifier first, and then the recognition results from each classifier are fused in the end. The fusion-based method at the decision level, without increasing the dimensionality, can combine various modalities by exploring the contributions of different emotional expressions. The error weighted classifier combination method (EWC) [59], [60] is a notable example. However, facial and vocal features are generally

TABLE II
VOCAL EXPRESSION-BASED EMOTION RECOGNITION

References	Feature	Approaches	Class	Evaluated Database	Accuracy
Jeon et al. [50]	• 384 acoustic features	• SVM for segment emotion model • Majority vote, GMM, and average of segment probabilities for later decision fusion	• Four emotion categories	• USC database: - one actress - 121 sentences with 4 different emotions • EMO-DB [53]: 339 sentences were selected	• Recognition rate for USC database: Unit type - (1) whole utterance: 84.9 - (2) 3 words: 87.0 - (3) phrases: 83.1 - (4) time based: 88.4 ((2), (3), (4) were decision by GMM) • Recognition rate for EMO-DB: Unit type - (1) whole utterance: 80.5 - (2) time based: 84.7 with GMM decision
Wu et al. [45]	• 253 acoustic-prosodic features	• GMM, SVM, MLP with/without the ESS	• Four emotion categories	• The corpora, totally contains 2,033 sentences collected by NCKU	• Recognition rate: - GMM without/with ESS: 68.73/72.61 - MLP without/with ESS: 69.86/71.87 - SVM without/with ESS: 75.33/78.16
Bitouk et al. [51]	• In total, the combined set consists of 261 spectral and prosodic features	• SVM	• Six emotion categories	• LDC [54] - 7 actors - 548 utterances • EMO-DB	• Classification rate for combined feature sets: - LDC dataset: 43.7 - EMO-DB: 78.2
Ntalampiras et al. [52]	• Mel filterbank • Pitch • wavelet domain features, etc	• The feature sets are integrated based on - (1) statistical - (2) spectral - (3) modeling using autoregressive processes • Classifier: - HMM - MLP - random forest, etc.	• Six emotion categories	• EMO-DB	• Feature sets integration methodology: Best result for PWP integration-spectral moments: - recall rate 63.5 - precision rate: 67.5 • Fusion at different level: Feature without/with temporal integration (HMM): - recall rate: 63.2/64.5 - precision rate : 64.4/66.7 • Log-likelihood with optimally integrated feature sets (the method using simple logistic): - recall rate: 92.2 - precision rate: 93.4, etc.
Lin et al. [48]	• Prosodic features: - pitch - energy - formant	• Temporal course modeling • HMM • SVM	• Four emotion categories • Two dimensional categories	• MHMC conversation-based affective speech corpus • 1,114 utterances	• Recognition rate for four emotion categories: - SVM: 50.22 - HMM: 56.50 - temporal course modeling: 79.82 • Recognition rate for dimensional categories: - SVM: 85.2 - HMM: 89.69 - Temporal course modeling: 89.69

MLP: Multilayer Perception, ESS: Emotional Salient Segment, PWP: Perceptual Wavelet Packet.

complementary to each other in emotional expression [1]. The assumption of conditional independence among multiple modalities at the decision level is inappropriate. Correlation between visual and audio modalities should be considered. To manage this problem, model-level fusion approaches [17], [23], [24], [61], [62] were proposed to exploit the information of correlation among multiple modalities and explore the temporal relationship between the visual and audio signal streams. Notable examples include Coupled Hidden Markov Model (C-HMM) [23], [24], Triple HMM (T-HMM) [61], Semi-Coupled HMM (SC-HMM) [23], [24] and Multi-stream Fused HMM (MF-HMM) [17]. Besides, a part of existing model-level fusion strategies gradually explored the evolution patterns of emotional expression in a conversational environment which considers the sub-emotional/emotional state transitions within/between sentences in a conversation. This approach not only considers the correlation between audio and visual streams but also explores sub-emotion/emotion evolution patterns [24], [62]. Furthermore, a more sophisticated fusion strategy called hybrid approach was recently proposed to combine feature-level and decision-level fusion or model-level and decision-level fusion strategies to obtain a better recognition result. A notable example is the Error Weighted Semi-Coupled HMM (EWSC-HMM) [23]. For EWSC-HMM, the state-based bimodal alignment strategy in SC-HMM is first proposed to align the temporal relation between audio and visual streams. The Bayesian classifier weighting scheme is then adopted to explore the contributions of the SC-HMM-based classifiers for different audio-visual feature pairs in order to obtain the optimal emotion recognition result. The illustration of the EWSC-HMM is shown in Fig. 5 and the formula of the final equation is shown in the following:

$$P(w \mid x^a, x^v) \approx \sum_{i=1}^{C}\sum_{j=1}^{D}\{\sum_{k=1}^{K} P(w \mid \tilde{w}_k, \lambda_i^a, \lambda_j^v) \\ [\max_{S^a, S^v} P(x^a, S^a \mid \lambda_i^a, \tilde{w}_k) P(S^v \mid S^a, \lambda^a, \tilde{w}_k) P(S^a \mid S^v, \lambda^v, \tilde{w}_k) \quad (2) \\ P(x^v, S^v \mid \lambda_j^v, \tilde{w}_k)] P(\tilde{w}_k \mid \lambda_i^a, \lambda_j^v)\} P(\lambda_i^a, \lambda_j^v \mid x^a, x^v)$$

TABLE III
FACIAL-VOCAL EXPRESSION-BASED EMOTION RECOGNITION

References	Fusion	Feature	Approaches	Class	Evaluated Database	Accuracy
Song et al. [61]	• M	• 18 FAPs • 48 prosodic features • 16 formant frequency features	• T-HMM	• Six emotion categories	• Collected from Institute of Automation, Chinese Academy of Sciences - Training: 100 samples for each emotion - Testing: overall emotion more than 500 samples	• Recognition rate for six emotion categories: higher than 91
Zeng et al. [17]	• M	• 12 MUs • prosodic features	• MF-HMM	• Seven emotion categories and four cognitive categories	• Collected from University of Illinois at Urbana-Champaign - 20 subjects - 660 sequences	• Recognition rate for eleven categories: 80.45
Wu et al. [24]	• F • D • M • H	• 30 FAPs • prosodic features	• FP • EWC • C-HMM • SC-HMM • EWSC-HMM • 2H-SC-HMM	• Four emotion categories for MHMC audio-visual database • Four emotion quadrants for SEMAINE database	• Posed: MHMC audio-visual database - 7 subjects - 1680 sentences • Spontaneous: SEMAINE - 4 subjects - 320 utterances	• Recognition rate for MHMC audio-visual database: - FP: 75.77 - EWC: 80.54 - C-HMM: 83.21 - SC-HMM: 85.24 - EWSC-HMM: 90.6 - 2H-SC-HMM: 91.55 • Recognition rate for SEMAINE database: - FP: 64.06 - EWC: 69.06 - C-HMM: 66.25 - SC-HMM: 72.19 - EWSC-HMM: 78.13 - 2H-SC-HMM: 87.5
Metallinou et al. [62]	• F • M • H	• Facial markers • Prosodic and spectral features	• HMM • C-HMM • HMM+HMM • HMM+BLSTM • BLSTM • RNN • BRNN • LSTM	• Dimensional space - Valence - Activation - 3 clusters - 4 clusters	• IEMOCAP [63]	• Comparing context free and context-sensitive classifiers based on F1-measure: • For Valence - BLSTM: 65.12 ± 5.13 • For Activation - HMM+HMM: 57.71 ± 4.23 • For 3 clusters - BLSTM: 72.35 ± 5.10 • For 4 clusters - BLSTM: 62.80 ± 6.69

Fusion: **F**eature/**D**ecision/**M**odel/**H**ybrid-level, FAPs: Facial Animation Parameters, MUs: Motion Units, FP: fusion facial and prosodic features at feature level, 2H-SC-HMM: two-level hierarchical alignment-based SC-HMM, BLSTM: Bidirectional Long Short-Term Memory neural network, RNN: Recurrent Neural Networks, BRNN: Bidirectional Recurrent Neural Networks, LSTM: Long Short-Term Memory neural network.

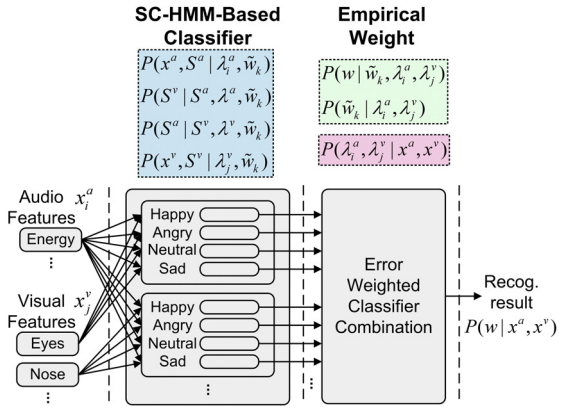

Fig. 5. The architecture of the error weighted semi-coupled HMMs.

where $P(w \mid \tilde{w}_k, \lambda_i^a, \lambda_j^v)$, $P(\tilde{w}_k \mid \lambda_i^a, \lambda_j^v)$, $P(\lambda_i^a, \lambda_j^v \mid x^a, x^v)$ represent the empirical weights and are used to weight the output of each SC-HMM-based classifier in order to obtain an optimally combined decision. The remaining terms denote the SC-HMM which contains the output probabilities for the audio and visual HMMs $P(x^a, S^a \mid \lambda_i^a, \tilde{w}_k)$ and $P(x^v, S^v \mid \lambda_j^v, \tilde{w}_k)$, and the state alignment probabilities $P(S^v \mid S^a, \lambda^a, \tilde{w}_k)$ and $P(S^a \mid S^v, \lambda^v, \tilde{w}_k)$.

Based on the analyses mentioned above, Table III provides an overview of the existing popular data fusion strategies for facial-vocal expression-based emotion recognition with respect to the utilized features, classifiers, and performances.

V. CONCLUSION

This paper provides a survey of current research work in audio-, visual-, and audiovisual-based emotion recognition. It

covers the theoretical background and applications ranging from salient emotional features to multi-modal data fusion strategies. Although a number of promising studies have been proposed and successfully applied to various applications, there are still some important issues yet to be addressed in the fields which include the following:

1. Toward spontaneous emotional expressions analysis, how to build a comprehensive (i.e., cover various factors such as personality trait, culture, age, language, etc.) and accessible benchmark database is indispensible to provide a consistent evaluation procedure.
2. Exploring the issues on the effects of a person's arbitrary movement, partial occlusion, speaking, complex temporal course of emotional expression, and background noise is beneficial to increase the system's value in real life applications.
3. A better automatic feature normalization approach should be considered and developed in the future, since most of the existing studies assume that the speaker ID are known, and have neutral examples (for normalization purpose) for each speaker beforehand.
4. Modeling the context information, that is, simultaneously considering the sub-emotional (i.e., onset, apex, and offset temporal phases) and emotional state transitions within and between utterances is useful in spontaneous emotional expression analysis.
5. Developing a better data fusion approach is desirable to obtain the advantages of various fusion strategies, such as combining the model-level and decision-level fusion properties.

REFERENCES

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

[2] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 31, no. 1, pp. 39-58, 2009.

[3] M. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.

[4] C. H. Wu, Z. J. Chuang, and Y. C. Lin, "Emotion recognition from text using semantic label and separable mixture model," *ACM Trans. on Asian Language Information Processing*, vol. 5, no. 2, pp. 165-183, Jun. 2006.

[5] S. Koelstra, M. Pantic, and I. Patras, "A dynamic texture-based approach to recognition of facial actions and their temporal models," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 32, no. 11, pp. 1940-1954, 2010.

[6] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions–the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 22, no. 12, pp. 1424-1445, 2000.

[7] M. Pantic and M. Bartlett, "Machine analysis of facial expressions," *Face Recognition*, K. Delac and M. Grgic, eds., pp. 377-416, I-Tech Education and Publishing, 2007.

[8] M. F. Valstar and M. Pantic, "Fully automatic recognition of the temporal phases of facial actions," *IEEE Trans. Systems, Man and Cybernetics–Part B*, vol. 42, no.1, pp. 28-43, 2012.

[9] P. Ekman and W. V. Friesen, *Picture of Facial Affect*. Palo Alto, Calif.: Consulting Psychologist, 1976.

[10] P. Ekman, "Facial expression and emotion," *A. Psychologist*, vol. 48, pp. 384-392, 1993.

[11] P. Ekman and W. Friesen, *The Facial Action Coding System: A Technique for The Measurement of Facial Movement*. Palo Alto, CA: Consulting Psychologists Press, 1978.

[12] P. Ekman, W. V. Friesen, and J. C. Hager, *Facial Action Coding System.* Salt Lake City, UT: A Human Face, 2002.

[13] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. Systems, Man and Cybernetics–Part B*, Accepted, 2013.

[14] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 33-80, 2001.

[15] F. Tian, Q. Zheng, R. Zhao, T. Chen, and X. Jia, "Can e-Learner's Emotion be Recognized from Interactive Chinese Texts?," *International Conference on Computer Supported Cooperative Work in Design*, pp. 546-551, 2009.

[16] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu, "Joint Emotion-Topic Modeling for Social Affective Text Mining," *IEEE International Conference on Data Mining*, pp. 699-704, 2009.

[17] Z. Zeng, J. Tu, B. M. Pianfetti, Jr., and T. S. Huang, "Audio-visual affective expression recognition through multistream fused HMM," *IEEE Trans. Multimedia*, vol. 10, no. 4, pp. 570-577, 2008.

[18] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161-1178, 1980.

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

[20] D. Wu, T. D. Parsons, E. Mower, and S. Narayanan, "Speech emotion estimation in 3D space," *Proceedings of IEEE International Conference on Multimedia & Expo (ICME)*, pp. 737-742, 2010.

[21] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68-99, 2010.

[22] M. A. Nicolaou, H. Gunes, and M. Pantic, "A multi-layer hybrid framework for dimensional emotion classification," *Proc. of ACM Multimedia*, pp. 933-936, 2011.

[23] J. C. Lin, C. H. Wu, and W. L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no.1, pp. 142-156, Feb. 2012.

[24] C. H. Wu, J. C. Lin, and W. L. Wei, "Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course," to appear in *IEEE Trans. on Multimedia*, DOI: 10.1109/TMM.2013.2269314, 2013.

[25] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 23, no. 6, pp. 681-685, 2001.

[26] G. Guo and C. R. Dyer, "Learning from examples in the small sample case: Face expression recognition," *IEEE Trans. Systems, Man and Cybernetics–Part B*, vol. 35, no. 3, pp. 477-488, 2005.

[27] K. Anderson and P. W. McOwan, "A real-time automated system for recognition of human facial expressions," *IEEE Trans. Systems, Man and Cybernetics–Part B*, vol. 36, no. 1, pp. 96-105, 2006.

[28] J. Whitehill and C. W. Omlin, "Haar features for FACS AU recognition," in *Proc. IEEE Int. Conf. Autom. Face and Gesture Recog*., pp. 217-222, 2006.
[29] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati, and J. F. Cohn, "Facial action unit recognition with sparse representation," in *Proc. IEEE Int. Conf. Autom. Face and Gesture Recog*., pp. 336-342, 2011.
[30] Y. Tong, W. Liao, and Q. Ji, "Facial action unit recognition by exploiting their dynamic and semantic relationships," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 29, no. 10, pp. 1683-1699, 2007.
[31] Y. Tong, J. Chen, and Q. Ji, "A unified probabilistic framework for spontaneous facial action modeling and understanding," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 32, no. 2, pp. 258-273, 2010.
[32] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Transactions on Image Processing*, vol. 22, no.7, pp. 2559-2573, 2013.
[33] O. Rudovic, M. Pantic, and I. Patras, "Coupled Gaussian processes for pose-invariant facial expression recognition," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 35, no. 6, pp. 1357-1369, 2013.
[34] C. H. Wu, W. L. Wei, J. C. Lin, and W. Y. Lee, "Speaking effect removal on emotion recognition from facial expressions based on eigenface conversion," to appear in *IEEE Trans. on Multimedia*, DOI: 10.1109/TMM.2013.2272917, 2013.
[35] J. C. Lin, C. H. Wu, and W. L. Wei, "Facial action unit prediction under partial occlusion based on error weighted cross-correlation model," to appear in *Int'l Conf. Acoustics, Speech, and Signal Processing*, 2013.
[36] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proc. IEEE Int. Conf. Autom. Face and Gesture Recog*., pp. 46-53, 2000.
[37] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of database," *Speech Comm.*, vol. 40, no. 1-2, pp. 33-60, 2003.
[38] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended Cohn-Kanade dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression," *Computer Vision and Pattern Recog*nition *Workshops*, pp. 94-101, 2010.
[39] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *Proceedings of IEEE International Conference on Multimedia & Expo (ICME)*, pp. 317-321, 2005.
[40] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Proc. IEEE Int. Conf. Autom. Face and Gesture Recog*., pp. 211-216, 2006.
[41] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proc. IEEE Int. Conf. Autom. Face and Gesture Recog*., pp. 1-8, 2008.
[42] G. Mckeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schroe, "The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no.1, pp. 5-17, 2012.
[43] C. H. Wu, J. F. Yeh, and Z. J. Chuang, "Emotion perception and recognition from speech," in *Affective Information Processing*. New York: Springer, ch. 6, pp. 93-110, 2009.
[44] S. G. Kooladugi, N. Kumar, and K. S. Rao, "Speech emotion recognition using segmental level prosodic analysis," *Int'l Conf. on Devices and Communications*, pp. 1-5, 2011.
[45] C. H. Wu and W. B. Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affective Computing*, vol. 2, no.1, pp. 1-12, 2011.
[46] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98-112, 2007.
[47] O. W. Kwon, K. Chan, J. Hao, and T. W. Lee, "Emotion recognition by speech signals," *Proc. Eighth European Conf. Speech Comm. and Technology*, 2003.
[48] J. C. Lin, C. H. Wu, and W. L. Wei, "Emotion recognition of conversational affective speech using temporal course modeling," to appear in *Proc. of the Interspeech*, 2013.
[49] B. Schuller and G. Rigoll, "Timing levels in segment-based speech emotion recognition," *Proc. of the Interspeech*, pp. 1818-1821, 2006.
[50] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," *Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 4940-4943, 2011.
[51] D. Bitouk, R. Verma, and A. Nenkova, "Class-level spectral features for emotion recognition," *Speech Communication*, vol. 52, no. 7-8, pp. 613-625, 2010.
[52] S. Ntalampiras and N. Fakotakis, "Modeling the temporal evolution of acoustic parameters for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 3, no.1, pp. 116-125, 2012.
[53] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *Proc. of the Interspeech*, pp. 1517-1520, 2005.
[54] Linguistic Data Consortium. Emotional prosody speech and transcripts. LDC Catalog No.: LDC2002S28, University of Pennsylvania, 2002.
[55] C. Busso et al., "Analysis of emotion recognition using facial expression, speech and multimodal information," *Proc. Sixth ACM Int'l Conf. Multimodal Interfaces*, pp. 205-211, 2004.
[56] B. Schuller, R. Muller, B. Hornler, A. Hothker, H. Konosu, and G. Rigoll, "Audiovisual recognition of spontaneous interest within conversations," *Proc. Ninth ACM Int'l Conf. Multimodal Interfaces*, pp. 30-37, 2007.
[57] Z. Zeng, Z. Zhang, B. Pianfetti, J. Tu, and T. S. Huang, "Audio-visual affect recognition in activation-evaluation space," *Proc. IEEE Int'l Conf. on Multimedia and Expo*, pp.828-831, 2005.
[58] Z. Zeng, J. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 424-428, Feb. 2007.
[59] A. Metallinou, S. Lee, and S. Narayanan, "Audio-visual emotion recognition using Gaussian mixture models for face and voice," *Proc. Int'l Symposium on Multimedia*, pp. 250-257, 2008.
[60] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," *Proc. 35th Int'l Conf. Acoustics, Speech, and Signal Processing*, pp. 2462-2465, 2010.
[61] M. Song, M. You, N. Li, and C. Chen, "A robust multimodal approach for emotion recognition," *Neurocomputing*, vol. 71, no. 10-12, pp. 1913-1920, 2008.
[62] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affective Computing*, vol. 3, no. 2, pp. 184-198, 2012.
[63] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.