Survey on audiovisual emotion recognition

Abstract

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.

Introduction

This paper gives a survey on the existing audiovisual emotion databases and recent advances in the research on audiovisual bimodal data fusion strategies.

AudioVisual emotion databases

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]
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]

1. Elicitation method
    1. posed (acted)

    2. induced (via clips)
    3. spontaneous (occurring during an interaction)
2. Emotion categorization
    1. discrete categorical representation
    2. continuous dimensional representation
    3. event representation (affective behavior; e.g. level of interest, depression, laughter, etc.)

AudioVisual bimodal fusion for emotion recognition

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) $49.99 \pm 3.63/61.92 \pm 4.88$ (V) $60.98 \pm 4.96/51.36 \pm 4.14$ (AV) $64.67 \pm 6.48/52.28 \pm 5.37$	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
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	(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

			couples of facial FP				
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 Fo, 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

1. Audio features

   Correlations among prosodic features and emotions

	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

   1. local (frame-level) features

      The local features represent the speech features extracted based on the unit of speech "frame".

      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)

   2. global (utterance-level) features

      The global features are calculated from the statistics of all speech features extracted from the entire "utterance"

      The set of functionals extracted from the LLDs, such as max, min, mean, standard deviation, duration, linear predictive coefficients (LPC)

   3. 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. features.

2. Facial Features

   1. Appearance

      Depict the facial texture such as wrinkles, bulges, and furrows

   2. Geometric

Represent the shape or location of facial components

3. Model

    1. Active appearance model

        The AAM achieved successful human face alignment, even for the human faces having non-rigid deformations.

    2. Local binary patterns

        Being the dense local appearance descriptors, local binary patterns (LBPs) have been used extensively for facial expression recognition in recent years

    3. The linear interpolation technique

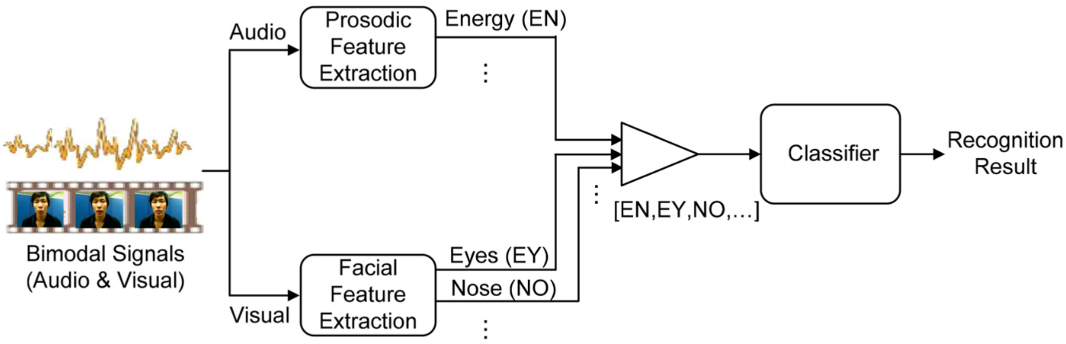
        For audiovisual data fusion, to deal with the problem of mismatched frame rates between audio and visual features, the linear interpolation technique has been widely applied, which interpolates the video features to match the frame rate of audio features

3. Bimodal fusion approaches

    1. Feature-level (early) fusion

        Facial and vocal features are concatenated to construct a joint feature vector, and are then modeled by a single classifier for emotion recognition

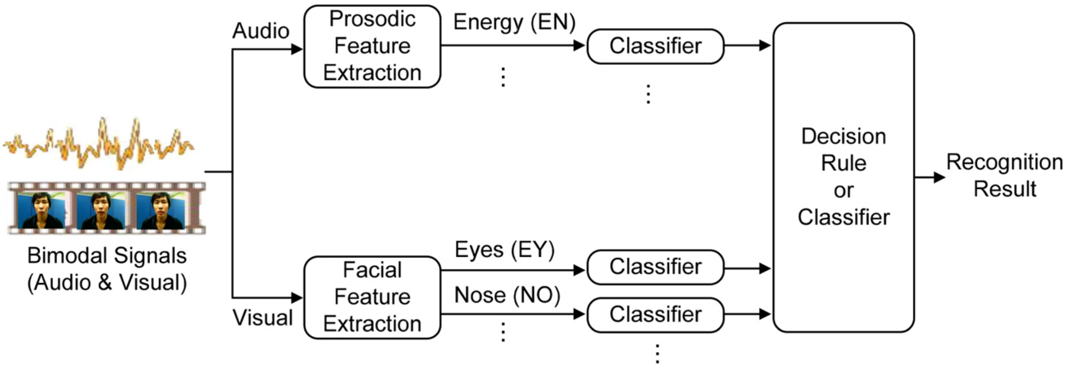
        Illustration of feature-level fusion strategy for audiovisual emotion recognition.



    2. The decision-level fusion

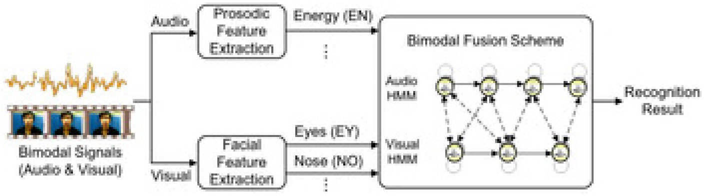
        Multiple signals can be modeled by the corresponding classifier first, and then the recognition results from each classifier are fused in the end, as shown in Fig. 5. The fusion-based method at the decision level, without increasing the dimensionality, can combine various modalities by exploring the contributions of different emotional expressions.

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



    3. A model-level fusion strategy

Be proposed to emphasize the correlation information between multimodalities and explore the temporal relationship between audio and video signal streams (as shown in Figure 6).



There are several distinctive examples such as Coupled HMM (C-HMM), Tripled HMM (T-HMM), Multistream Fused HMM (MFHMM), and Semi-Coupled HMM (SC-HMM)

4. Hybrid approach

   Be proposed to integrate different fusion approaches to obtain a better emotion recognition result.

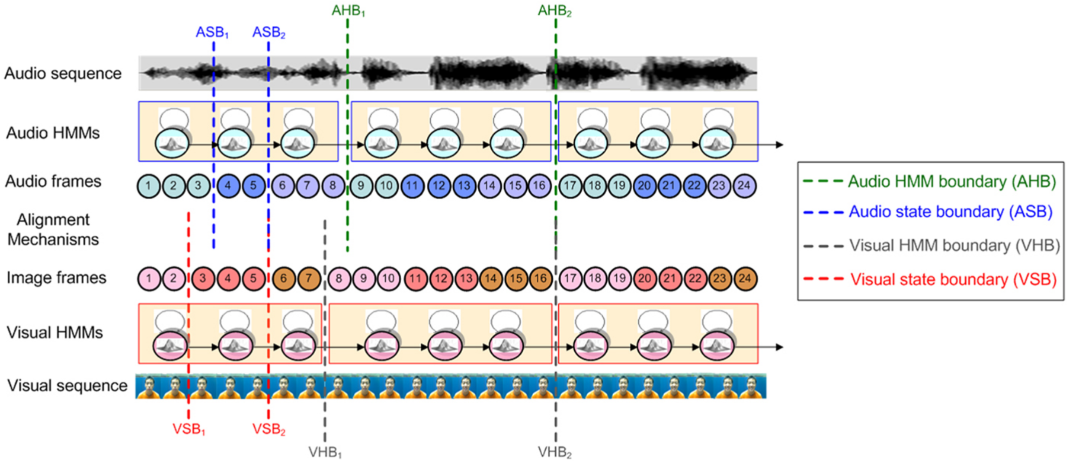
   The Error Weighted SC-HMM (EWSC-HMM), as an example of the hybrid approach, consists of model-level and decision-level fusion strategies and concurrently combines both advantages.

4. A few related issues

   1. Another important issue in audiovisual data fusion is related to the problem of asynchrony between audiovisual signals. For audiovisual data fusion, current feature-level fusion methods deal with asynchrony based on strict constraints on temporal synchronization between modalities or using static features from each input utterance (i.e., ignoring temporal information). Therefore, under the assumption of strict temporal synchronization, feature-level fusion does not work well if the input features of human voices and facial expressions differ in temporal features. Furthermore, since decision-level fusion methods focus on exploring how to effectively combine recognition outputs from separate audio and video classifiers that independently model audio and video signal flows, synchronization can be ignored in decision-level fusion problem. On the one hand, model-level fusion methods (such as C-HMM, T-HMM, SC-HMM, T_AsyDBN, etc.) Recently proposed and applied to audiovisual emotion recognition, it attempts to model asynchronous voice and facial expressions and maintain their natural correlation over time. Unlike the dynamic programming algorithms (Viterbi and forward analysis) used in traditional HMMs to deal with temporal changes, current model-level fusion methods Extended to handle synchronization issues by desynchronizing audio and video streams and aligning audiovisual signals at the state level. Therefore, current model-level fusion methods such as C-HMM can achieve good performance for audiovisual signals with large synchronization deviations.
   2. Furthermore, for naturalistic emotion recognition, several existing fusion strategies explore the evolutionary patterns of emotion expression in dialogue environments. These methods take into account intra-sentence/inter-sentence emotional sub-states or emotional state transitions in a dialogue, not only exploiting the correlation between audio and video streams, but also exploring the evolution patterns of emotional sub-states or emotional states. Previous studies have shown that a complete emotional expression can be divided into three consecutive temporal phases, onset (application), apex (release), and offset (relaxation), which take into account the mode and intensity of the expression.

      An example of the temporal phases of onset, apex, and offset of facial expression is shown in Fig. 7.

[Figure 8](#) illustrates model- and state-level alignments between audio and visual HMM sequences in the happy emotional state.



5. 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.