Chapter 5

Multimodal Affect Recognition: Current Approaches and Challenges

Hussein Al Osman and Tiago H. Falk

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/65683

Abstract

Many factors render multimodal affect recognition approaches appealing. First, humans employ a multimodal approach in emotion recognition. It is only fitting that machines, which attempt to reproduce elements of the human emotional intelligence, employ the same approach. Second, the combination of multiple-affective signals not only provides a richer collection of data but also helps alleviate the effects of uncertainty in the raw signals. Lastly, they potentially afford us the flexibility to classify emotions even when one or more source signals are not possible to retrieve. However, the multimodal approach presents challenges pertaining to the fusion of individual signals, dimensionality of the feature space, and incompatibility of collected signals in terms of time resolution and format. In this chapter, we explore the aforementioned challenges while presenting the latest scholarship on the topic. Hence, we first discuss the various modalities used in affect classification. Second, we explore the fusion of modalities. Third, we present publicly accessible multimodal datasets designed to expedite work on the topic by eliminating the laborious task of dataset collection. Fourth, we analyze representative works on the topic. Finally, we summarize the current challenges in the field and provide ideas for future research directions.

Keywords: affect recognition, multimodal, machine learning, sensor fusion

1. Introduction

Humans employ rich emotional communication channels during social interaction by modulating their speech utterances, facial expressions, and body gestures. They also rely on emotional cues to resolve the semantics of received messages. Interestingly, humans also

communicate emotional information when interacting with machines. They express affects and respond emotionally during human-machine interaction. However, machines, from the simplest to the most intelligent ones devised by humans, have conventionally been completely oblivious to emotional information. This reality is changing with the advent of affective computing.

Affective computing advocates the idea of emotionally intelligent machines. Hence, these machines can recognize and simulate emotions. In fact, over the last decade, we have witnessed a steadily increasing interest in the development of automated methods for human-affect estimation. The applications of such technologies are varied and span several domains. Rosalind Picard, in her 1997 book Affective Computing, describes various applications, such as a computer tutor that personalizes learning based on the user's affective response, affective agent that assists autistic individuals navigate difficult social situations, and a classroom barometer that informs the teacher of the level of engagement of the students [1]. Numerous other applications have been proposed over the years. For instance, many researchers suggest the creation of emotionally intelligent computers to improve the quality of the human-computer interaction (HCI) [2–4]. Other affective computing applications abound in the literature. For example, Gilleade et al. [5] propose the use of affective methods in video gaming. Al Osman et al. [6] present a mobile application for stress management. However, regardless of the application, all researchers in the field are faced with the following questions: How can a machine classify human emotions? What should the machine do in response to the recognized emotions? In this chapter, we are solely concerned with the first question.

Various strategies of affect classification have been successfully employed under restricted circumstances. The primary modalities that have been thoroughly explored pertain to facial-expression estimation, speech-prosody (tone) analysis, physiological signal interpretation, and body-gesture examination. In this chapter, we explore affect-recognition techniques that integrate multiple modalities of affect expression. These techniques are known in the literature as multimodal methods.

Although, today, most of the affective computing applications are unimodal, the multimodal approach has been advocated by numerous researchers [4, 7–14]. There are many reasons that render the multimodal approach appealing. First, humans employ a multimodal approach in emotion recognition. It is only fitting that machines, which attempt to reproduce elements of human emotional intelligence, employ the same approach. Second, the combination of multiple-affective signals not only provides a richer collection of data but also helps alleviate the effects of uncertainty in the raw signals. After all, these signals are collected by imperfect sensors with numerous possible sources of error between the signal producer and processor. Lastly, it potentially gives us the flexibility to classify emotions even when one or more source signals are not possible to retrieve. This can happen in situations where the face or body is partially or fully occluded, which disqualifies the visual modality, or when the user is not speaking which eliminates the vocal modality from consideration. However, the multimodal approach presents challenges pertaining to the fusion of individual signals,

dimensionality of the feature space, and incompatibility of collected signals in terms of time resolution and format.

Before we proceed, we clarify a potential source of confusion. The terms affect and emotion can have different meanings in various fields. For instance, according to Shouse, a researcher in communication, an emotion refers to the display of a feeling, whether it is genuine or feigned [15]. However, an "affect is a non-conscious experience of intensity" [15]. Some psychologists consider affect as the experience of emotion [16]. In this chapter, we consider the terms emotion and affect to be synonymous since a sizable amount of works in affective computing use them interchangeably.

The remainder of this chapter is organized as follows: Section 2 summarizes the modalities of affect recognition, Section 3 describes pertinent modality-fusion techniques, Section 4 presents publicly available multimodal emotional databases, Section 5 surveys representative multimodal affect-recognition methods, and Section 6 discusses the challenges in the field and future research directions.

2. Modalities of affect recognition

In this section, we explore the various modalities of emotional channels that can be used for the automated resolution of human affect. The fundamental question that this section addresses is the following: What measurable information the machine needs to retrieve and interpret to estimate human affect?

When it comes to judging expressive behaviors, humans rely in general on verbal and nonverbal channels [17]. The verbal channels correspond to speech, while nonverbal channels include the eye gaze and blink, facial and body expression, and speech prosody. Note that speech corresponds to the semantics of the communicated message while speech prosody is concerned with the tonal content of voice regardless of the meaning of spoken phrases. Facial expression and speech prosody are believed to be the most relied upon by humans for emotions' interpretation [18]. Hence, these channels are likely rich in informational cues about the affective state. Social psychologists have interestingly remarked that expressive behaviors can be consciously regulated to convey a calculated self-presentation. However, nonverbal channels tend to be less vulnerable to deliberate manipulation. Moreover, when verbal behavior conflicts with nonverbal comportment, nonverbal expressions may be more reflective of the true affective status [17]. In fact, researchers have found speech prosody to be the least consciously controllable modality [19]. The latter finding can inform the development of affective applications for lie detection. In the following subsections, we detail the commonly used modalities of affect recognition.

2.1. Visual modalities

The visual modality is rich in relevant informational content and includes the facial expression, eye gaze, pupil diameter, and blinking behavior, and body expression. We explore these affective sources in this section.

2.1.1. Facial expression

The most studied nonverbal affect-recognition method is facial-expression analysis [20]. Perhaps, that is because facial expressions are the most intuitive indicators of affect. Even as children, we draw simplistic faces that convey various emotions by manipulating the forehead creases, eyebrows, and mouth. We also find it instinctive to use emoticons in digital textual communications that convey emotions through simple facial-expression depictions.

2.1.1.1. Facial muscle movement coding

Facial expressions result from the contraction of facial muscles resulting in the temporary deformation of the neutral expression. These deformations are typically brief and last mostly between 250 ms and 5 s [21]. Darwin [22] is one of the early researchers to explore the evolutionary foundation of facial-expressions display. He argues that facial expressions are universal across humans. He contends that they are habitual movements associated with certain states of the mind. These habits have been favored through natural selection and inherited across generations. Ekman and Fiesen [23] built on the idea of facial-expression universality to conceive the facial action coding system (FACS) that describes all possible perceivable facial muscle movements in terms of predefined action units (AUs). All AUs are numerically coded and facial expressions correspond to one or more AUs. Although FACS is primarily employed to detect emotions, it can be used to describe facial muscle activation regardless of the underlying cause. Inspired by FACS, other facial expression coding systems have been proposed, such as the emotional facial action coding system (EMFACS) [24], the maximally descriptive facial movement coding system (MAX) [25], and the system for identifying affect expressions by holistic judgment AFFEX [26]. The latter systems are solely directed at emotion recognition.

The Moving Pictures Experts Group (MPEG) defined the facial animation parameters (FAPs) in the MPEG-4 standard to enable the animation of face models. MPEG-4 describes facial feature points (FPs) that are controlled by FAPs. The value of the FAP corresponds to the magnitude of deformation of the facial model in comparison to the neutral state. Though the standard was not originally intended for automated emotion detection, it has been employed for that goal in various works [27, 28]. These coding systems inspired researchers to develop automated image or video-processing methods that track the movement of facial features to resolve the affective state [29].

2.1.1.2. Facial-expression detection

Facial-expression detection algorithms involve the following three steps: (1) face detection (or face tracking across video frames), (2) feature extraction, and (3) affect classification. We will not discuss face detection or tracking in this chapter, the reader can refer to the plethora of existing literature on the topic (e.g., [30–32]).

Feature extraction is an essential aspect of expression recognition. Jiang et al. [33] divide the feature extraction methods into two types: geometric-based and appearance-based methods. Geometric features typically correspond to the distances between key facial points or

the velocity vectors of these points as the facial expression develops. However, appearance features reflect the changes in image texture resulting from the deformation of the neutral expression (e.g., facial bulges and creases) [33]. We detail few feature extraction schemes employed across many works. Each technique listed represents a set of methods that apply the same basic idea in feature extraction:

- Motion estimators: They are geometric-based feature extraction methods. They estimate the motion between two images. The most commonly used algorithm is optical flow [34]. When the latter is used for facial feature extraction, the camera is usually assumed to be stationary and the nonrigid motion resulting from facial deformation is tracked across video frames. The output is a series of vectors that represent motion. This technique has been used in numerous works, either alone [35–37], or in combination with other feature extraction techniques [38].
- Point trackers: They are geometric-based feature extraction methods. They track feature points across an image sequence. A typical algorithm, known as the Kanade-Lucas-Tomasi (KLT) tracker [39, 40], computes the spatial translation or affine transformation of features between consecutive video frames. Spatiotemporal vectors can be obtained from the movement of tracked features.
- Gabor wavelets: They are appearance-based feature extraction methods. They typically use a set of Gabor filters at different scales and orientation for feature extraction. Gabor filters are a type of band-pass filters that act in a similar manner to the human cortical cells by mostly resolving edges of objects present in an image. This technique usually involves training a machine-learning model using Gabor features extracted from a database of facial expression and running the model to classify emotions from images.

For classification, numerous techniques have been proposed such as support vector machine (SVM), neural network (NN), and hidden Markov models (HMMs) [29, 35, 41–45].

In addition to facial-expression analysis, eye-based features such as pupil diameter, gaze distance, and gaze coordinates, and blinking behavior have been used in multimodal systems [10, 12]. In fact, Panning et al. [10] found that in their multimodal system, the speech paralinguistic features and eye-blinking frequency were the most contributing modalities to the classification process.

2.1.2. Body expression

The importance of body expressions for affect recognition has been debated in the literature, with conflicting opinions. McNeill [46] maintains that two-handed gestures are closely associated with the spoken verbs. Hence, they arguably do not present new affective information; they simply accompany the speech modality. Consequently, some researchers argue that gestures may play a secondary role in the human recognition of emotions [4, 13]. This suggests that they might be less reliable than other modalities in delivering affective cues that can be automatically analyzed. However, increasingly, there is more evidence toward the viability of this method in affect recognition, at least for a subset of affective expressions [20, 47–51].

In fact, Lhommet and Marsella [52] contend that body expressions are harder to control consciously than facial expressions, and therefore might reflect more genuine emotions.

Affect recognition using body expression involves tracking the motion of body features in space. Many works rely on the use of three-dimensional (3D) measurement systems that require markers to be attached to the subject's body [11, 53–56]. However, some markerless solutions involving video cameras [57, 58] and wearable sensors [59] have been proposed. Once the motion is captured, a variety of features are extracted from body movement. In particular, the following features have been reliably used: velocity of the body or body part [11, 53, 55, 60–64], acceleration of the body or body part [11, 55, 60, 61, 64], amount of movement [11, 64], joint positions [62], nature of movement (e.g., contraction, expansion, and upward movement) [11], orientation of body parts (e.g., head and shoulder) [54, 56, 63, 64], and angle or distance between body parts (e.g., distance from hand to shoulder and angle between shoulder-shoulder vectors) [54, 56, 61, 63]. Using these features, a variety of classification models have been suggested, such as decision tree [11], multilayered perceptron (MLP) [53, 59], SVM [55, 61, 63], naïve Bayes [63], and HMM [62].

2.2. Audio modality

Speech carries two interrelated informational channels: linguistic information that express the semantics of the message and implicit paralinguistic information conveyed through prosody. Both of these channels carry affective information. Hence, in this section, we briefly describe the general mechanisms of extracting affect from these channels.

2.2.1. Linguistic speech channel

Humans often explain how they feel during social interaction. Hence, building an understanding of the spoken message provides a straightforward way of assessing affect. This technique of affect recognition falls under the wider topic of sentiment analysis and opinion mining using natural language processing. Typically, an automatic speech recognition algorithm is used to convert speech into a textual message. Then, a sentiment analysis method interprets the polarity or emotional content of the message. However, this approach for affect recognition has its pitfalls. First, it is not universal, and therefore a natural language speech processor has to be developed for each dialect; second, it is vulnerable to masking since humans are not always forthcoming about their emotional status [17].

In this section, we only discuss sentiment analysis. We will not cover automatic speech recognition. The readers can consult the survey of Benzeghiba et al. [65] for a thorough treatment of this topic. Sentiment analysis methods can broadly be divided into two categories: lexicon-based techniques and statistical-learning approaches. Lexicon-based techniques classify affect based on the presence of unambiguous affect words or phrases in the text. Numeric values are tied to these words or phrases. Hence, overall sentiment can be extracted through a scoring system that results from the aggregation of these values. Statistical-learning methods, in turn, generate a bag of words whose elements are used as features in machine-learning algorithms. Hybrid approaches that propose a combination of these techniques have also been studied [66, 67].

http://dx.doi.org/10.5772/65683

2.2.2. Paralinguistic speech-prosody channel

Sometimes, it is not about what we say, but how we say it. Therefore, speech-prosody analyzers ignore the meaning of messages and focus on acoustic cues that reflect emotions. Before the extraction of tonal features from speech, preprocessing is often necessary to enhance, denoise, and dereverberate the source signal [68]. Then, using windowing functions, low-level descriptor (LLDs) features are extracted at usually 100 frames per second with segment sizes between 10 and 30 ms. Windowing functions are usually rectangular for time-domain features and smooth for frequency or time-frequency features. Numerous LLDs can be extracted, and we list a few: pitch (fundamental frequency F_0), energy (e.g., maximum, minimum, and root mean square), linear prediction cepstral (LPC) coefficients, perceptual linear prediction coefficients, cepstral coefficients (e.g., mel-frequency cepstral coefficients, MFCCs), formants (e.g., amplitude, position, and width), and spectrum (mel-frequency and FFT bands) [68–72]. Linguistic LLDs can also be retrieved, such as word and phoneme sequences [68, 69]. Recently, speech-modulation spectral features were also shown to contain complementary information to prosodic and cepstral features [73].

For classification, global statistics features are classified using static classifier such as SVM [69, 74–76]. Short-term features are processed though dynamic classifiers, such as HMM [68, 76]. Due to the large number of possible features, researchers have proposed the use of dimension-reduction schemes such as principal component analysis (PCA) [69] or linear discriminant analysis (LDA) [68]. More recently, with the burgeoning of deep-learning principles, deep neural networks have also been explored for speech emotion recognition, with very promising results (e.g., [77–79]).

2.3. Physiological modality

Physiological signals can be used for affect recognition through the detection of biological patterns that are reflective of emotional expressions. These signals are collected through typically noninvasive sensors that are affixed to the body of the subject. However, brain imaging [80] and remote physiological monitoring schemes [81, 82] have been proposed.

There are a multitude of physiological signals that can be analyzed for affect detection. Typical physiological signals used for the assessment of affect are electrocardiography (ECG), electromyography (EMG), electroencephalograph (EEG), skin conductance (also known as galvanic skin response, and electrodermal activity), respiration rate, and skin temperature. ECG records the electrical activity of the heart. Conventionally, 12 electrodes are connected to various parts of the body to conduct this measurement. However, in affective computing, most systems use the Lead I configuration that requires only two electrodes [6]. From the ECG signal, the heart rate (HR) and heart rate variability (HRV) can be extracted. HRV is used in numerous studies that assess mental stress [6, 83–85]. EMG measures muscle activity and is known to reflect negatively valenced emotions [86]. EEG is the electrical activity of the brain measured through electrodes connected to the scalp and possibly forehead. There is little agreement on the number of electrodes to use or features to extract from EEG. EEG features are often used to classify emotional dimensions of arousal [87–90], valence [88–90], and dominance [90, 91]. Skin conductance measures the resistance of the skin by passing a negligible

current through the body. The resulting signal is reflective of arousal [86] as it corresponds to the activity of the sweat glands. The latter are controlled by the autonomous nervous system (ANS) that regulates the flight or fight response. Finally, respiration rate tends to reflect arousal [92], while skin temperature carries valence cues [93].

3. Multimodal fusion techniques

With multimodal affect-recognition approaches, information extracted from each modality must be reconciled to obtain a single-affect classification result. This is known as multimodal fusion. The literature on this topic is rich and generally describes three types of fusion mechanisms: feature-level fusion, decision-level fusion, and hybrid approaches. In this section, we present the general principles behind these techniques and describe key ideas related to each type.

3.1. Feature-level fusion

A common method to perform modality fusion is to create a single set from all collected features. A single classifier is then trained on the feature set. This method is advocated by Pantic et al. [4, 13] as it mimics the human mechanism of tightly integrating information collected through various sensory channels. However, feature-level fusion is plagued by several challenges. First, the larger multimodal feature set contains more information than the unimodal one. This can present difficulties if the training dataset is limited. Hughes [94] has proven that the increase in the feature set may decrease classification accuracy if the training set is not large enough. Second, features from various modalities are collected at different time scales [13]. For example, frequency domain HRV features typically summarize seconds or minutes' worth of data [6], while speech features can be in the order of milliseconds [13]. Third, a large feature set undoubtedly increases the computational load of the classification algorithm [95]. Finally, one of the advantages of multimodal affect recognition is the ability to produce an emotion classification result in the presence of missing or corrupted data. However, feature-level fusion is more vulnerable to the latter issues than decision-level fusion techniques [96].

3.2. Decision-level fusion

Typically, a classifier makes errors in some area of the feature space [97]. Hence, combining the results of multiple classifiers can alleviate this shortcoming. This is especially true when each classifier is operating on a different modality that corresponds to a separate feature space.

Using decision-level fusion, modalities can be independently classified using separate models and the results are joined using a multitude of possible methods. Therefore, this approach is said to employ an ensemble of classifiers. Ensemble members can belong to the same family or different families of statistical classifiers. In fact, static and dynamic classifiers can both be employed in such a multimodal system.

3.2.1. Combination strategies based on voting

The simplest and one of the oldest methods to achieve decision-level fusion is to use a voting mechanism [98]. Hence, the classification reached by the majority of the ensemble members is

adopted as the outcome. However, a tie in the votes can be reached if the number of classifiers is odd. This disqualifies bimodal affect-recognition systems. Furthermore, even for an odd number of classifiers, a definite decision cannot be guaranteed if more than two classes are being considered [95] (e.g., the six prototypical emotions). The classification of a single affect is a typical binary problem that can be solved using this approach. A system that monitors a single affect such as stress or frustration can use this approach as long as an odd number of modalities are supported.

3.2.2. Combination strategies based on prior knowledge

In many cases, it is crucial to assess the performance of each classifier to inform decision making during the combination process. For instance, using the training dataset, we can calculate the confusion matrix for each classifier. Given an ensemble of C classifiers, the confusion matrix of classifier c_i, where $i = 1..C$, is described by

$$Pc_i = \begin{bmatrix} n_{11}^i & \cdots & n_{1M}^i \\ \vdots & \ddots & \vdots \\ n_{M1}^i & \cdots & n_{MM}^i \end{bmatrix} \tag{1}$$

where njki corresponds to the number of times c_i classified an observed sample x as belonging to class r_j while in reality it belongs to class r_k, and M is the total number of classes. The diagonal of the confusion matrix where $j = k$ represents the times where the classifier was correct.

To overcome the limitations of the voting approach, a weighted majority voting scheme can be used. In this approach, classifiers are not treated as equal peers and their votes are weighted to reduce the probability of a tie. The weights can be calculated based on the performance of the classifier in terms of recognition and error rates retrieved from the confusion matrix during training or using a test dataset after training [95, 98, 99]. Lam and Suen [99] propose an optimization process that uses a genetic algorithm to compute the voting weights. They observe that there is often a trade-off between recognition, rejection, and error rates. Therefore, they attempt to maximize objective function (1):

$$F = \text{recognition} - \beta \times \text{error} \tag{2}$$

where β is a constant that can take on different values depending on the accuracy and reliability desired [99]. Hence, in the genetic algorithm, F is used as the fitness value.

Beyond the use of voting schemes, Huang and Suen [100] use a lookup table during training to keep track of the combinations of classifier outputs along with the correct class and number of occurrence of this combination. The number of occurrence reflects the confidence level that the corresponding combination produces the recorded correct class. When the latter combination is observed, the outcome with the highest confidence level, as recorded in the lookup table, is chosen. Gupta et al., in turn, proposed a quality-aware decision fusion scheme, where classifiers were developed for several physiological modalities (i.e., EEG, ECG, GSR, and facial features) and their individual decisions were weighted by the measured quality of each raw signal [101]. Experimental results showed that system failure rates due to noisy segments were drastically reduced, and improved affect-recognition performance could be achieved [101].

Kim and Lingenfelser [102] introduce an ensemble combination strategy that accounts for the capability of some ensemble members to classify certain classes better than others. Therefore, they rank the classes according to the accuracy of their classification across all ensemble members using the confusion matrices produced from the training data. To reach an ensemble decision for an observed sample, the classifier corresponding to the highest-ranked class performs the classification. We refer to that class as the test class. If the classification result matches the test class, then that result is taken to be the ensemble decision. If not, then the next class in the ranked list becomes the test class and the procedure is repeated. If we do not obtain a match for any of the classes, then the classifier with the best overall performance on the training data is tasked with the classification on behalf of the ensemble.

Lastly, Gupta, Laghari, and Falk have made use of a variant of the SVM called relevance vector machines (RVMs) for affect recognition. RVMs have the same functional form of SVMs but are embedded into a Bayesian framework [103]. Therefore, for classification, RVMs compute the probabilities of class membership rather than the point estimates. These class membership probabilities can be seen as a measure of classifier "confidence" and were used as weights for decision-level fusion [90]. While the work in [90] focuses only on a single modality, EEG, it fused the decisions of classifiers trained on different classes of EEG features (power spectral, asymmetry, and graph theoretic), and thus the observed advantages could also be seen for multimodal setups.

3.2.3. Combination strategies for continuous output classifiers

For the ensemble decision of continuous output problems, the probabilities for each class over all classifiers can be used for fusion. Lingenfelser et al. [95] refer to this probability as support and we adopt this terminology. Using these probabilities, several decision-level combination rules are conceived. We detail only a subset of these rules. The maximum rule stipulates that the ensemble decision for an observed feature vector corresponds to the class with the largest support. The sum rule sums the total support for each class chosen by any of the classifiers. Then, the class with the largest support is chosen as the ensemble decision. Similarly, the mean rule calculates the mean support for each chosen class as opposed to the sum. Instead of calculating the mean, a weighted average of total support for each chosen class can also be calculated. Finally, the product rule is similar to the sum rule, except for the use of the multiplication operation instead of the addition for the calculation of the total support.

3.3. Hybrid fusion

When a fusion technique combines feature and decision-level fusion, it is referred to as a hybrid-fusion scheme. For instance, we can achieve fusion in two stages. In the first stage, a classifier can perform feature-level fusion. For example, a single classifier can handle features from audio and video signals. In the second stage, decision-level fusion can be used to combine the results of that classifier with another one operating on physiological (e.g., HRV) features.

Ref. [104] proposes a simple hybrid-fusion approach where the result from the feature-level fusion is fed as an additional input to the decision-level fusion stage. Lingenfelser et al. [95]

propose two variants of one method called the one versus rest. This approach creates an ensemble composed of classifiers trained on each feature set (i.e., features from a modality). However, these classifiers model a two-class problem. That is, each one of them is specialized in classifying a single class. One last multiclass classifier is added to the ensemble and is trained on the merged feature set (i.e., features from all modalities). For the first variant, during classification, for an observed sample, the support for a class obtained from its two-class classifiers is multiplied with the support of the multiclass classifier to obtain an accumulated support. The class with the highest accumulated support is chosen as the ensemble decision. The second variant is similar, except that it chooses the best two-class classifier for each class and uses it to calculate accumulated support.

3.4. Dimensionality problem

Affective information tends to be highly dimensional. It is not unusual for a feature set to contain thousands of variables. Valstar and Pantic [105] model the facial action temporal dynamics by extracting 2520 features from each facial video frame. The problem can be further exasperated when multiple modalities are considered. Feature-level fusion techniques are especially vulnerable to this problem. For instance, Kim and Lingenfelser [102] extract 1280 speech and 26 physiological features to classify affect. Two strategies are generally adopted to reduce the feature space dimension. First, feature-selection techniques that choose a subset of the feature set for model construction are widely used [7, 12, 28, 104]. Second, dimension-reduction methods such as principal component analysis and linear discriminant analysis are commonly employed [7, 10, 106].

4. Multimodal datasets

One of the challenges in developing multimodal affect-recognition methods is the need to collect multisensory data from a large number of subjects. Also, it is difficult to compare the obtained results with other studies given that the experimental setup varies. Therefore, it is essential to use databases to streamline research efforts on the topic and produce repeatable and easy-to-compare results. Very few multimodal affect databases are publicly available. We divide these databases into three types: posed, induced, and natural-emotional databases. For the posed databases, the subjects are asked to act out a specific emotion while the result is captured. Typically, facial and body expression and speech information are captured in posed databases. However, posed databases have their limitations, as they cannot incorporate biosignals; it cannot be guaranteed that posed emotions trigger the same physiological response as spontaneous ones [107]. For the induced databases, the subjects are exposed to a stimulus (e.g., watching a video) in a controlled setting, such as laboratory. The stimulus is designed to evoke certain emotions. In some cases, following the stimulus, the subjects are explicitly asked to act out an emotional expression. The eNTERFACE'05 [108] is an example of such database. These databases combine aspects of induced and posed emotions. For the natural databases, the subjects are exposed to a real-life stimulus such as interaction with human or machine. Data collection mostly occurs in a noncontrolled environment. The AFEW database

[109] presents annotated video clips from movies. Therefore, although the emotional expressions are acted out by professional actors, they take place in real-world environments (or at least simulated ones). Since these expressions are likely to be as subtle as naturally occurring ones, as actors strive to mimic realistic behavior, we categorize this database as a natural one. We concede that it does not perfectly fit in any of the three presented types.

For the induced and natural databases, the measured sensory information is labeled with the emotional information. The label is usually obtained through subject self-assessment, observer/listener judgment, or FACS coding (manually coded facial expressions). Self-assessment is performed using tools such as self-assessment Manikin (SAM) [110] or feeltrace [111]. **Table 1** shows a list of publicly accessible multimodal emotional databases. Most of the databases address the visual and audio modalities, while few recent ones introduce physiological channels.

Reference	DB type	# Subjects	Modalities	Affects	Labeling
GEMEP (2012) [112]	Posed	10	Visual and audio	Amusement, pride, joy, relief, interest, pleasure, hot anger, panic fear, despair, irritation, anxiety, sadness, admiration, tenderness, disgust, contempt, and surprise	N/A
SAL (2008) [113]	Induced	24	Visual and audio	Dimensional and categorical labeling	Feeltrace
Belfast (2000) [114]	Natural	24	Visual and audio	Dimensional and categorical labeling	Feeltrace
MIT (2005) [83]	Natural	17	Physiological (ECG, EMG, skin conductance, and respiration)	Low, medium, and high stress	Observers' judgment
HUMAINE (2007) [115]	Induced and natural	Multiple databases	Visual, audio, and physiological (ECG, skin conductance and temperature, and respiration)	Varies across databases	Observers' judgment + self-assessment
VAM (2008) [116]	Natural	19	Visual and audio	Dimensional labeling	SAM
SEMAINE (2010) [117]	Induced	20	Visual and audio	Dimensional labeling and six basic emotions	Observers' judgment
DEAP (2012) [118]	Induced	32	Visual for (22 subjects) and physiological (EEG, ECG, EMG, and skin conductance)	Dimensional labeling	SAM
MAHNOB-HCI (2012) [12]	Induced	27	Visual (face + eye gaze), audio, and physiological (EEG, ECG, skin conductance and temperature, and respiration)	Dimensional and categorical labeling	Self-assessment (SAM for arousal and valence)

Reference	DB type	# Subjects	Modalities	Affects	Labeling
eNTERFACE'05 (2006) [108]	Posed + induced	42	Visual and audio	Six basic emotions	Observers' verification
RECOLA (2013) [119]	Natural	46	Visual, audio, and physiological (ECG and skin conductance)	Dimensional labeling	Observers' judgment
PhySyQX (2015) [120]	Natural	21	Audio and physiological (EEG and near-infrared spectroscopy, NIRS)	Dimensional labeling	SAM (valence, arousal, dominance) plus nine other quality metrics (e.g., naturalness, acceptance)
AFEW (2012) [109]	Natural	N/A(1426 video clips)	Visual and audio	Six basic emotions + neutral	Expressive keywords from movie subtitles + observers' verification

Table 1. Summary of the characteristics of publicly accessible multimodal emotional databases.

5. Multimodal affect detection

Humans display emotions through a variety of behaviors that are difficult for a machine to fully appreciate. They modulate their facial muscles, eye gaze, body gestures, gait, and speech tone among other channels of expression to convey emotions. Therefore, the understanding of these emotional cues requires a multisensory system that is able to track several or all of these channels.

Many multimodal affect-recognition schemes have been proposed. They generally differ in terms of the modalities, classification method, and fusion mechanism used, and emotions recognized. In **Table 2**, we survey several representative multimodal affect-recognition studies. Facial-expression analysis features prominently in these studies, followed by speech prosody. However, there seems to be little agreement on the nature and number of the features to be extracted for each modality.

All of the reviewed works consider a subset of possible features that can be extracted from the dataset. Therefore, effective feature selection is required to simplify the classification models, and reduce training time and overfitting. Hence, diverse automated techniques are employed for that purpose, such as the wrapper method [28], analysis of variance (ANOVA)-based approach [12], sequential backward selection [7], minimum redundancy maximum relevance [121], and correlation-based feature selection [104]. Some works rely on expert knowledge [27, 106] as an effective feature-selection scheme. Furthermore, several works elect to reduce the dimensionality of the feature space using PCA [7, 10, 106].

Reference	Modalities	Classifier**	Features	Affects	DB type	Overall recognition rate*
Castellano et al. [28]	Visual (face, body) and audio	BN	**Face:** statistical values from FAPs and their derivatives **Body:** quantity of motion and contraction index of the body, velocity, acceleration, and fluidity of the hand's barycenter **Speech:** intensity, pitch, MFCC, Bark spectral bands, voiced segment characteristics, and pause length (377 features in total)	Anger, despair, interest, pleasure, sadness, irritation, joy and pride	Posed	FLF: 78.3% DLF: 74.6%
Panning et al. [10]	Visual (face and body) and audio	PCA+MLP	**Face:** eye blink per minute, mouth deformations, eyebrow actions **Body:** touch hand to face (binary) **Speech:** 36 features (12 MFCCs, their deltas and accelerations, and the zero-mean coefficient)	Frustration	Natural	FLF: 40–90%
Busso et al. [7]	Visual (face) and audio	SVM	**Face:** Four-dimensional feature vectors **Speech:** mean, standard deviation, range, maximum, minimum, and median of pitch and intensity	Anger, sadness, happiness, neutral	Posed	FLF: 89.1% DLF: 89.0%
Kapoor et al. [123]	Visual (face, posture) and physiological	GP	**Face:** nod and shakes, eye blinks, mouth activities, shape of eyes and eyebrows **Posture:** pressure matrices (on chair while seated) **Physiological:** skin conductance **Behavioral:** pressure on mouse	Frustration	Natural	FLF: 79%
Soleymani et al. [12]	Physiological + eye gaze	SVM (RBF Kernel)	**Physiological:** 20 GSR, 63 ECG, 14 respiration, 4 skin temperature, and 216 EEG features **Eye gaze:** pupil diameter, gaze distance, gaze coordinates	Arousal and valence	Induced	DLF: 72%

http://dx.doi.org/10.5772/65683

Reference	Modalities	Classifier**	Features	Affects	DB type	Overall recognition rate*
Kapoor and Picard [9]	Visual (face, and posture) and context	MGP	**Face:** Five features from upper face and two features from lower face **Posture:** current posture and level of activity **Context:** level of difficulty, state of the game	Student interest level	Natural	FLF: 86%
Paleari et al. [14]	Visual (face) and audio	NN	**Face:** 24 features corresponding to 12 pairs of feature points + 14 distance features **Speech:** 26 features, *F*0, formants (*F*1–*F*3), energy, harmonicity, LPC1 to LPC9, MFCC1 to MFCC10)	Six basic emotions	Induced + posed	DLF: 75%
Kim et al. [104]	Audio and physiological	LDF	**Physiological:** EMG at the nape of the neck, ECG, skin conductance, and respiration (26 features in total) **Speech:** pitch, utterance, energy, and 12 MFCC features	Positive/high, positive/low, negative/high, and negative/low	Induced	DLF: 57% FLF: 66% HF: 60%
Lin et al. [27]	Visual (face) and audio	C– HMM, SC-HMM, and EWSC-HMM	**Face:** FAPs calculated from 68 feature points on eyebrows, eyes, nose, mouth, and facial contour **Speech:** pitch, energy, and formants (*F*1–*F*5)	Joy, anger, sadness, and neutral	Posed	FLF: 75% DLF: 80% HF: 83–91%
				Valence and arousal quadrants	Induced	FLF: 64% DLF:69% HF: 66–78%
Ringeval et al. [106]	Visual (face), audio, and physiological	SVR + NN	**Face:** 84 appearance based features (after PCA based reduction) obtained from local Gabor binary patterns from three orthogonal planes + 196 geometric features based on 49 tracked facial landmarks **Speech:** One energy, 25 spectral (e.g., MFCC, spectral flux), and 16 voicing (e.g., F0, formants, and jitter) features **Physiological:** ECG (HR + HRV) and skin conductance	Valence and arousal	Natural	DLF: average correlation with self-assessment of 42%
Gupta et al. [101]	Visual (face/ head-pose) and physiological	SVM, NB	**Face/Head-pose:** lips thickness, spatial ratios (e.g., upper to lower lip thickness, eye brows to lips width) **Physiological:** ECG (power spectral features over ECG and HRV), skin conductance (power spectral, zero-crossing rate, rise time, fall time), EEG (band powers for δ-, θ-, α-, β-, and γ-bands)	Valence, arousal, and liking of multimedia content	Natural	DLF: *F*1-score of 59% (SVM) and 57% (NB)

Reference	Modalities	Classifier**	Features	Affects	DB type	Overall recognition rate*
Kaya and Salah [121]	Visual (face) and audio	ELM	**Face:** image is divided into 16 regions. 177 dimensional descriptors are extracted from each region using a local binary pattern histogram **Audio:** 1582 features such as *F*0, MFCC (0–14), and line spectral frequencies (0–7)	Six basic emotions + neutral	Natural	DLF: 44.23%

*FLF: Feature-Level Fusion, DLF: Decision-Level Fusion, HF: Hybrid Fusion.
**HMM: Hidden Markov Mode, C-HMM: Coupled HMM, SC-HMM: Semi-Coupled HMM, EWSC-HMM: Error Weighted SC-HMM, SVR: Support Vector Regression, LDF: Linear Discrimination Function, NN: Neural Networks, GP: Gaussian Process, MGP: Mixture of Gaussian Processes, MLP: Multilayer Perceptron, BN: Bayesian Network, NB: Naïve Bayes. ELM: Extreme Learning Machine.

Table 2. Representative multimodal affect-recognition studies.

Three modality-fusion techniques are commonly employed. There seems to be somewhat conflicting results concerning the most effective class of modality-fusion methods. For instance, Kapoor and Picard [9] obtain better results using feature-level fusion. Conversely, Busso et al. [7] fail to realize a discernible difference between the two methods. Beyond the latter two approaches, Lin et al. [27] propose three hybrid approaches that use coupled HMM, semi-coupled HMM, and error-weighted semi-coupled HMM based on a Bayesian classifier-weighing method. Their results show improvements over feature-and decision-level fusion for posed and induced-emotional databases. However, Kim et al. [104] were not able to improve over decision-level fusion with their proposed hybrid approach. The presence of confounding variables such as modalities, emotions, classification technique, feature selection and reduction approaches, and datasets used limits the value of comparing fusion results across studies. Consequently, Lingenfelser et al. [95] conducted a systematic study of several feature-level, decision-level, and hybrid-fusion techniques for multimodal affect detection. They were not able to find clear advantages for one technique over another.

Various affect classification methods are employed. For dynamic classification where the evolving nature of an observed phenomenon is classified, HMM is the prevalent choice of classifier [27]. For static classification, researchers use a variety of classifiers and we were not able to discern any clear advantages of one over another. However, an empirical study of unimodal affect recognition through physiological features found an advantage for SVM over *k*-nearest neighbor, regression tree, and Bayesian network [122]. Yet, a systematic investigation of the effectiveness of classifiers for multimodal affect recognition is needed to address the issue.

The database type seems to have an effect on the overall affect-recognition rate. We notice that studies that use posed databases generally achieve higher levels of accuracy compared to ones that use other types (e.g., [7, 27]). In fact, Lin et al. [27] perform an analysis of recognition rates using the same methods on two database types: posed and induced. They achieve significantly better results with the posed database. Natural databases result in typically lower recognition rates (e.g., [10, 101, 106, 121]) with the exception of studies [9, 123] that classify a single affect.

6. Discussion and conclusion

In this chapter, we have reviewed and presented the various affect-detection modalities, multimodal affect-recognition schemes, modality-fusion methods, and public multimodal-emotional databases. Although the work on multimodal human-affect classification has been ongoing for years, there are still many challenges to overcome. In this section, we detail these challenges and describe future research directions.

6.1. Current challenges

Numerous studies found multimodal methods to perform as good as or better than unimodal ones [9, 14, 27, 28, 104, 106]. However, the improvements of multimodal systems over unimodal ones are modest when affect detection is performed on spontaneous expressions in natural settings [124]. Also, multimodal methods introduce new challenges that have not been fully resolved. We summarize these challenges as follows:

- Multimodal affect-recognition methods require multisensory systems to collect the relevant data. These systems are more complex than unimodal ones in terms of the number and diversity of sensors involved and the computational complexity of the data-interpreting algorithms. This challenge is more evident when data are collected in a natural setting where user movement is not constrained to a controlled environment. Most physiological sensors are wearable and sensitive to movement. Therefore, additional signal filtering and preparation are required. Audio and visual data quality depends heavily on the distance between the subject and sensors and the presence of occluding objects between them.

- Multimodal affect-recognition methods necessitate the fusion of the modal features extracted from the raw signals. It is still unclear which fusion techniques outperform the others [95]. It seems that the performance of the fusion technique depends on the number of modalities, features extracted, types of classifiers, and the dataset used in the analysis [95]. While the first steps toward a quality-aware fusion system have been proposed [101], more research is still needed in order to gauge the true benefit of such an approach.

- It is still not understood what type and number of modalities are needed to achieve the highest level of accuracy in affect classification. Also, it is unclear how each modality contributes to the effectiveness of the system. Very few studies attempt to test the effect of single modalities on the overall performance [10] and a systematic study of the issue is still required.

- It is well established that context affects how humans express emotions [125, 126]. Nonetheless, context is disregarded by most work on affect recognition [127]. Therefore, we still need to address the challenge of incorporating contextual information into the affect classification process. Some attempts have been done in this regard [9, 123, 128–131]. For instance, Kim [128] suggests a two-stage procedure, where in the first stage, the affective dimensions of valence and arousal are classified, and in the second stage, the uncertainties between adjacent emotions in the two dimensional-affective space are resolved using

contextual information. However, more work is needed to validate this method and propose other similar methods that incorporate a rich set of contextual features.

- Although we have had major improvements in terms of the availability of public multimodal affect datasets over the past few years, many of the works in the area still use private datasets [127]. The use of nonpublic datasets makes results across studies challenging to compare and progress in the field difficult to trace.
- Multimodal-affective systems collect potentially private information such as video and physiological data. Special care needs to be afforded to the protection of such sensitive data. To the best of our knowledge, no work has specifically addressed this issue yet in the context of affective computing.
- In addition to the abundant technical challenges, the ethical implications of designing emotionally intelligent machines and how this can affect the human perception of these machines must be queried.

Despite these challenges, the results achieved in the last decade are very encouraging and the community of researchers on the topic is growing [124].

6.2. Future research directions

Several streams of research are still worth pursuing in the domain. For instance, more investigation is required on the usefulness and applicability of fusion techniques to different modalities and feature sets. Existing studies did not find consistent improvement in the accuracy of affect recognition between feature- and decision-level fusion. However, decision-level fusion schemes are advantageous when it comes to dealing with missing data [96]. After all, multisensory signal collection systems are prone to lost or corrupted segments of data. The introduction of effective hybrid-fusion techniques can further improve accuracy of classification. An empirical and exhaustive study of classifiers in multimodal emotion detection systems is still needed to gain a better understanding about their effectiveness. Although we have seen a flurry of new multimodal emotional databases in the last few years, there is still a need to create richer databases with larger amounts of data and support for more modalities. Moreover, new sensors and wearable technologies are emerging continuously, which may open doors for new affect-recognition modalities. For example, functional near-infrared spectroscopy (fNIRS) has been recently explored within this context [132]. fNIRS, much like functional magnetic resonance imagining (fMRI), measures cerebral blood flow and hemoglobin concentrations in the cortex, but at a fraction of the cost, without the interference of MRI acoustic noise, and with the advantage of being portable. Moreover, recent studies have explored the extraction of physiological information (e.g., heart rate and breathing) from face videos [81, 82], and thus may open doors for multimodal systems, which, in essence, would require only one modality (i.e., video). Notwithstanding, the biggest research challenge that remains is the detection of natural emotions. We have seen in this chapter that the accuracy of detection method decreases when natural emotions are classified. This is mainly due to the subtlety of the natural emotions (compared to exaggerated posed ones) and their dependence on the context [126]. Therefore, we expect that a considerable amount of future research will be dedicated for this effort.

http://dx.doi.org/10.5772/65683

Author details

Hussein Al Osman[1] and Tiago H. Falk[2]*

*Address all correspondence to: falk@emt.inrs.ca

1 University of Ottawa, Ottawa, Ontario, Canada

2 Institut National de la Recherche Scientifique, INRS-EMT, University of Quebec, Montreal, Quebec, Canada

References

[1] R. W. Picard, *Affective computing*. Cambridge, MA: MIT Press, 1997.

[2] R. W. Picard, "Affective computing for HCI," in *HCI,* vol. 1, pp. 829–833, 1999.

[3] T. Partala and V. Surakka, "The effects of affective interventions in human–computer interaction," *Interacting with Computers,* vol. 16, pp. 295–309, 2004.

[4] M. Pantic, N. Sebe, J. F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proceedings of the 13th annual ACM international conference on multimedia*, 2005, pp. 669–676.

[5] K. Gilleade, A. Dix, and J. Allanson, "Affective videogames and modes of affective gaming: assist me, challenge me, emote me," *Proceedings of DiGRA*, 2005.

[6] H. Al Osman, H. Dong, and A. El Saddik, "Ubiquitous biofeedback serious game for stress management," *IEEE Access,* vol. 4, pp. 1274–1286, 2016.

[7] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, *et al.*, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on multimodal interfaces*, 2004, pp. 205–211.

[8] Z. Zeng, Y. Hu, Y. Fu, T. S. Huang, G. I. Roisman, and Z. Wen, "Audio-visual emotion recognition in adult attachment interview," in *Proceedings of the 8th international conference on multimodal interfaces*, 2006, pp. 139–145.

[9] A. Kapoor and R. W. Picard, "Multimodal affect recognition in learning environments," in *Proceedings of the 13th annual ACM international conference on multimedia*, 2005, pp. 677–682.

[10] A. Panning, I. Siegert, A. Al-Hamadi, A. Wendemuth, D. Rösner, J. Frommer, *et al.*, "Multimodal affect recognition in spontaneous hci environment," in *2012 IEEE international conference on signal processing, communication and computing (ICSPCC),* 2012, pp. 430–435.

[11] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe, "Multimodal analysis of expressive gesture in music and dance performances," in *International gesture workshop*, 2003, pp. 20–39.

[12] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing,* vol. 3, pp. 42–55, 2012.

[13] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE,* vol. 91, pp. 1370–1390, 2003.

[14] M. Paleari, B. Huet, and R. Chellali, "Towards multimodal emotion recognition: a new approach," in *Proceedings of the ACM international conference on image and video retrieval*, 2010, pp. 174–181.

[15] E. Shouse, "Feeling emotion affect", *Media Culture Journal*, vol. 8, no. 6, pp. 1, 2005.

[16] M. A. Hogg and D. Abrams, "Social cognition and attitudes," in Martin, G. Neil and Carlson, Neil R. and Buskist, William, eds. *Psychology*, third edition, Pearson Education Limited, 2007, pp. 684–721.

[17] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin,* vol. 111, p. 256, 1992.

[18] N. Fragopanagos and J. G. Taylor, "Emotion recognition in human–computer interaction," *Neural Networks,* vol. 18, pp. 389–405, 2005.

[19] R. Rosenthal and B. M. DePaulo, "Sex differences in accommodation in nonverbal communication," In *Skill in nonverbal communication: Individual differences*, Cambridge, MA: Oelgeschlager, Gunn and Hain, 1979, pp. 68–103.

[20] B. de Gelder, "Why bodies? Twelve reasons for including bodily expressions in affective neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 364, pp. 3475–3484, 2009.

[21] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognition,* vol. 36, pp. 259–275, 2003.

[22] C. Darwin, *The expression of the emotions in man and animals*, London, UK: John Murray, 1965.

[23] P. Ekman and W. V. Friesen, "*Facial action coding system*," Palo Alto: Consulting Psychologists Press, 1978.

[24] W. V. Friesen and P. Ekman, "*EMFACS-7: Emotional facial action coding system*," Unpublished manuscript, San Francisco: University of California, 1983.

[25] C. E. Izard, "*The maximally discriminative facial movement coding system*", Newark, Canada: Academic Computing Services and University Media Services, University of Delaware, revised edition, 1983.

[26] C. E. Izard, L. M. Dougherty, and E. A. Hembree, *"A system for identifying affect expressions by holistic judgments (AFFEX)*", Instructional Resources Center, University of Delaware, 1983.

[27] J.-C. Lin, C.-H. Wu, and W.-L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Transactions on Multimedia,* vol. 14, pp. 142–156, 2012.

[28] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and emotion in human-computer interaction*, Berlin Heidelberg, Germany: Springer, 2008, pp. 92–103.

[29] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding,* vol. 91, pp. 160–187, 2003.

[30] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain, "Face detection in color images," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, pp. 696–706, 2002.

[31] C. Zhang and Z. Zhang, "*A survey of recent advances in face detection*," ed: Tech. rep., Microsoft Research, 2010.

[32] E. Hjelmås and B. K. Low, "Face detection: A survey," *Computer Vision and Image Understanding,* vol. 83, pp. 236–274, 2001.

[33] B. Jiang, M. F. Valstar, and M. Pantic, "Action unit detection using sparse appearance descriptors in space-time video volumes," in *2011 IEEE international conference on automatic face & gesture recognition and workshops (FG 2011)*, 2011, pp. 314–321.

[34] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Computing Surveys (CSUR),* vol. 27, pp. 433–466, 1995.

[35] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 36, pp. 96–105, 2006.

[36] Y. Yacoob and L. S. Davis, "Recognizing human facial expressions from long image sequences using optical flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 18, pp. 636–642, 1996.

[37] M. Kenji, "Recognition of facial expression from optical flow," *IEICE Transactions on Information and Systems,* vol. 74, pp. 3474–3483, 1991.

[38] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 23, pp. 97–115, 2001.

[39] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *IJCAI*, 1981, pp. 674–679.

[40] J. Shi and C. Tomasi, "Good features to track," in *1994 IEEE computer society conference on computer vision and pattern recognition, 1994. Proceedings CVPR'94*, 1994, pp. 593–600.

[41] M. Pardàs and A. Bonafonte, "Facial animation parameters extraction and expression recognition using Hidden Markov Models," *Signal Processing: Image Communication,* vol. 17, pp. 675–688, 2002.

[42] P. Michel and R. El Kaliouby, "Real time facial expression recognition in video using support vector machines," in *Proceedings of the 5th international conference on multimodal interfaces*, 2003, pp. 258–264.

[43] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions on Image Processing,* vol. 16, pp. 172–187, 2007.

[44] I. Kotsia, I. Buciu, and I. Pitas, "An analysis of facial expression recognition under partial facial image occlusion," *Image and Vision Computing,* vol. 26, pp. 1052–1067, 2008.

[45] M.S. Bartlett, G. Littlewort, I. Fasel, R. Movellan, "Real time face detection and facial expression recognition: Development and application to human computer interaction", *Proc. CVPR workshop on computer vision and pattern recognition for human-computer interaction*, vol. 5, 2006.

[46] D. McNeill, *Hand and mind: What gestures reveal about thought*, Chicago, IL: University of Chicago Press, 1992.

[47] P. E. Bull, *Posture & gesture*, Oxford, England: Pergamon Press, 1987.

[48] M. Argyle, *Bodily communication* (2nd ed.), London, England: Methuen, 1988.

[49] L. McClenney and R. Neiss, "Posthypnotic suggestion: A method for the study of nonverbal communication," *Journal of Nonverbal Behavior,* vol. 13, pp. 37–45, 1989.

[50] H. K. Meeren, C. C. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, pp. 16518–16523, 2005.

[51] J. Van den Stock, R. Righart, and B. De Gelder, "Body expressions influence recognition of emotions in the face and voice," *Emotion,* vol. 7, p. 487, 2007.

[52] Lhommet M., Marsella S.C., "Expressing emotion through posture," In Calvo R., D'Mello S., Gratch J., Kappas A., *The Oxford handbook of affective computing*, Oxford, UK: Oxford University Press, 2014, pp. 273–285.

[53] F. E. Pollick, V. Lestou, J. Ryu, and S.-B. Cho, "Estimating the efficiency of recognizing gender and affect from biological motion," *Vision Research,* vol. 42, pp. 2345–2355, 2002.

[54] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *International conference on affective computing and intelligent interaction*, 2007, pp. 48–58.

[55] L. Gong, T. Wang, C. Wang, F. Liu, F. Zhang, and X. Yu, "Recognizing affect from non-stylized body motion using shape of Gaussian descriptors," in *Proceedings of the 2010 ACM symposium on applied computing*, 2010, pp. 1203–1206.

[56] N. Bianchi-Berthouze and A. Kleinsmith, "A categorical approach to affective gesture recognition," *Connection Science,* vol. 15, pp. 259–269, 2003.

[57] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *2011 6th ACM/IEEE international conference on human-robot interaction (HRI)*, 2011, pp. 305–311.

[58] K. Vermun, M. Senapaty, A. Sankhla, P. Patnaik, and A. Routray, "Gesture-based affective and cognitive states recognition using kinect for effective feedback during e-learning," in *2013 IEEE fifth international conference on technology for education (T4E)*, 2013, pp. 107–110.

[59] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 41, pp. 1027–1038, 2011.

[60] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen, "Gesture-based affective computing on motion capture data," in *International conference on affective computing and intelligent interaction*, 2005, pp. 1–7.

[61] D. Bernhardt and P. Robinson, "Detecting affect from non-stylised body motions," in *International conference on affective computing and intelligent interaction*, 2007, pp. 59–70.

[62] D. Bernhardt and P. Robinson, "Detecting emotions from connected action sequences," in *International visual informatics conference*, 2009, pp. 1–11.

[63] M. Karg, K. Kuhnlenz, and M. Buss, "Recognition of affect based on gait patterns," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 40, pp. 1050–1061, 2010.

[64] N. Savva, A. Scarinzi, and N. Bianchi-Berthouze, "Continuous recognition of player's affective body expression as dynamic quality of aesthetic experience," *IEEE Transactions on Computational Intelligence and AI in games,* vol. 4, pp. 199–212, 2012.

[65] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication,* vol. 49, pp. 763–786, 2007.

[66] Y. Xia, E. Cambria, A. Hussain, and H. Zhao, "Word polarity disambiguation using bayesian model and opinion-level features," *Cognitive Computation,* vol. 7, pp. 369–380, 2015.

[67] J. M. Chenlo and D. E. Losada, "An empirical study of sentence features for subjectivity and polarity classification," *Information Sciences,* vol. 280, pp. 275–288, 2014.

[68] F. Weninger, M. Wöllmer, and B. Schuller, "Emotion recognition in naturalistic speech and language—a survey," in *Emotion Recognition: A Pattern Analysis Approach*, Hoboken, NJ: John Wiley & Sons, Inc., 2015, pp. 237–267.

[69] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, et al., "The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2253–2256.

[70] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology,* vol. 70, p. 614, 1996.

[71] C. Jones and J. Sutherland, "Acoustic emotion recognition for affective computer gaming," in *Affect and emotion in human-computer interaction*, Berlin Heidelberg, Germany: Springer, 2008, pp. 209–219.

[72] F. Eyben, M. Wöllmer, and B. Schuller, "OpenEAR—introducing the Munich open-source emotion and affect recognition toolkit," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*, 2009, pp. 1–6.

[73] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication,* vol. 53, pp. 768–785, 2011.

[74] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, *et al.*, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing,* vol. 27, pp. 1760–1774, 2009.

[75] B. Schuller and G. Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features," in *Proc. INTERSPEECH*, Brighton, UK, 2009, pp. 1999–2002.

[76] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation," in *Affect and emotion in human-computer interaction*, Berlin Heidelberg, Germany: Springer, 2008, pp. 75–91.

[77] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2011, pp. 5688–5691.

[78] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 511–516.

[79] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, *et al.*, "Hybrid deep neural network-Hidden Markov Model (DNN-HMM) based speech emotion recognition," in *2013 humaine association conference on affective computing and intelligent interaction (ACII)*, 2013, pp. 312–317.

[80] T. Dalgleish, B. D. Dunn, and D. Mobbs, "Affective neuroscience: Past, present, and future," *Emotion Review,* vol. 1, pp. 355–368, 2009.

[81] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering,* vol. 58, pp. 7–11, 2011.

[82] D. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five band digital camera," *IEEE Transactions on Biomedical Engineering,* vol. 61, pp. 2593–2601, 2014.

[83] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems,* vol. 6, pp. 156–166, 2005.

[84] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard, "The effect of mental stress on heart rate variability and blood pressure during computer work," *European Journal of Applied Physiology,* vol. 92, pp. 84–89, 2004.

[85] E. Jovanov, A. D. Lords, D. Raskovic, P. G. Cox, R. Adhami, and F. Andrasik, "Stress monitoring using a distributed wireless intelligent sensor system," *IEEE Engineering in Medicine and Biology Magazine,* vol. 22, pp. 49–55, 2003.

[86] A. Nakasone, H. Prendinger, and M. Ishizuka, "Emotion recognition from electromyography and skin conductance," in *Proc. of the 5th international workshop on biosignal interpretation*, 2005, pp. 219–222.

[87] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *International workshop on multimedia content representation, classification and security*, 2006, pp. 530–537.

[88] Z. Khalili and M. Moradi, "Emotion detection using brain and peripheral signals," in *2008 Cairo international biomedical engineering conference*, 2008, pp. 1–4.

[89] R. Horlings, D. Datcu, and L. J. Rothkrantz, "Emotion recognition using brain activity," in *Proceedings of the 9th international conference on computer systems and technologies and workshop for PhD students in computing*, 2008, p. 6.

[90] R. Gupta and T. H. Falk, "Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization," *Neurocomputing,* vol. 174, pp. 875–884, 2016.

[91] A. Clerico, R. Gupta, and T. H. Falk, "Mutual information between inter-hemispheric EEG spectro-temporal patterns: A new feature for automated affect recognition," in *2015 7th international IEEE/EMBS conference on neural engineering (NER)*, 2015, pp. 914–917.

[92] I. Homma and Y. Masaoka, "Breathing rhythms and emotions," *Experimental Physiology,* vol. 93, pp. 1011–1021, 2008.

[93] S. E. Rimm-Kaufman and J. Kagan, "The psychological significance of changes in skin temperature," *Motivation and Emotion,* vol. 20, pp. 63–78, 1996.

[94] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory,* vol. 14, pp. 55–63, 1968.

[95] F. Lingenfelser, J. Wagner, and E. André, "A systematic discussion of fusion techniques for multi-modal affect recognition tasks," in *Proceedings of the 13th international conference on multimodal interfaces*, 2011, pp. 19–26.

[96] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Transactions on Affective Computing,* vol. 2, pp. 206–218, 2011.

[97] L. A. Alexandre, A. C. Campilho, and M. Kamel, "On combining classifiers using sum and product rules," *Pattern Recognition Letters,* vol. 22, pp. 1283–1289, 2001.

[98] C. Y. Suen and L. Lam, "Multiple classifier combination methodologies for different output levels," in *International workshop on multiple classifier systems*, 2000, pp. 52–66.

[99] L. Lam and C. Y. Suen, "Optimal combinations of pattern classifiers," *Pattern Recognition Letters,* vol. 16, pp. 945–954, 1995.

[100] Y. S. Huang and C. Y. Suen, "The behavior-knowledge space method for combination of multiple classifiers," in *IEEE computer society conference on computer vision and pattern recognition*, 1993, pp. 347–347.

[101] R. Gupta, M. Khomami Abadi, J. A. Cárdenes Cabré, F. Morreale, T. H. Falk, and N. Sebs, "A quality adaptive multimodal affect recognition system for user-centric multimedia indexing," in *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, 2016, pp. 317–320.

[102] J. Kim and F. Lingenfelser, "Ensemble approaches to parametric decision fusion for bimodal emotion recognition," in *Proc. BIOSIGNALS*, Valencia, Spain, 2010, pp. 460–463.

[103] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research,* vol. 1, pp. 211–244, 2001.

[104] J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner, "Integrating information from speech and physiological signals to achieve emotional sensitivity," in *Proc. INTERSPEECH*, Lisboa, Portugal, 2005, pp. 809–812.

[105] M. F. Valstar and M. Pantic, "Combined support vector machines and hidden markov models for modeling facial action temporal dynamics," in *International workshop on human-computer interaction*, 2007, pp. 118–127.

[106] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne*, et al.*, "Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th international workshop on audio/visual emotion challenge*, 2015, pp. 3–8.

[107] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th international colloquium on signal processing and its applications (CSPA)*, 2011, pp. 410–415.

[108] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *22nd international conference on data engineering workshops (ICDEW'06)*, 2006, pp. 8–8.

[109] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, pp. 34–31, 2012.

[110] J. D. Morris, "Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response," *Journal of Advertising Research,* vol. 35, pp. 63–68, 1995.

[111] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[112] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the Geneva multimodal expression corpus for experimental research on emotion perception," *Emotion,* vol. 12, p. 1161, 2012.

[113] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. Heylen, "The sensitive artificial listner: An induction technique for generating emotionally coloured conversation," 2008.

[114] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: Considerations, sources and scope," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[115] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, *et al.*, "The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data," in *International conference on affective computing and intelligent interaction*, 2007, pp. 488–500.

[116] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *2008 IEEE international conference on multimedia and expo*, 2008, pp. 865–868.

[117] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *2010 IEEE international conference on multimedia and expo (ICME)*, 2010, pp. 1079–1084.

[118] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, *et al.*, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing,* vol. 3, pp. 18–31, 2012.

[119] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, 2013, pp. 1–8.

[120] R. Gupta, H. J. Banville, and T. H. Falk, "PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience," in *2015 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)*, 2015, pp. 1–5.

[121] H. Kaya and A. A. Salah, "Combining modality-specific extreme learning machines for emotion recognition in the wild," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 487–493.

[122] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis and Applications,* vol. 9, pp. 58–69, 2006.

[123] A. Kapoor, W. Burleson, and R. W. Picard, "Automatic prediction of frustration," *International journal of human-computer studies,* vol. 65, pp. 724–736, 2007.

[124] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Computing Surveys (CSUR),* vol. 47, p. 43, 2015.

[125] C. E. Izard, "Innate and universal facial expressions: evidence from developmental and cross-cultural research," 1994.

[126] U. Hess, R. Banse, and A. Kappas, "The intensity of facial expression is determined by underlying affective state and social situation," *Journal of personality and social psychology,* vol. 69, p. 280, 1995.

[127] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence,* vol. 31, pp. 39–58, 2009.

[128] J. Kim, "Bimodal emotion recognition using speech and physiological changes", *Robust Speech Recognition and Understanding*, pp. 265–280, 2007.

[129] K. Forbes-Riley and D. J. Litman, "Predicting emotion in spoken dialogue from multiple knowledge sources," in *HLT-NAACL*, 2004, pp. 201–208.

[130] D. J. Litman and K. Forbes-Riley, "Predicting student emotions in computer-human tutoring dialogues," in *Proceedings of the 42nd annual meeting on association for computational linguistics*, 2004, p. 351.

[131] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing,* vol. 3, pp. 184–198, 2012.

[132] R. Gupta, S. Arndt, J.-N. Antons, S. Möllery, and T. H. Falk, "Characterization of human emotions and preferences for text-to-speech systems using multimodal neuroimaging methods," in *2014 IEEE 27th Canadian conference on electrical and computer engineering (CCECE)*, 2014, pp. 1–5.