Multimodal Technologies and Interaction

MDPI

Review

A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database

Mohammad Faridul Haque Siddiqui [1], Parashar Dhakal [2], Xiaoli Yang [3] and Ahmad Y. Javaid [4,*]

[1] Department of Computer Science, West Texas A&M University, Canyon, TX 79016, USA; msiddiqui@wtamu.edu
[2] Manufacturing Department, Grote Industries, Madison, IN 47250, USA; parashar.dhakal@grote.com
[3] Department of Computer Science and Engineering, Fairfield University, Fairfield, CT 06824, USA; xyang@fairfield.edu
[4] Electrical Engineering and Computer Science Department, The University of Toledo, Toledo, OH 43606, USA
* Correspondence: ahmad.javaid@utoledo.edu; Tel.: +1-419-530-8191

Abstract: Multimodal human–computer interaction (HCI) systems pledge a more human–human-like interaction between machines and humans. Their prowess in emanating an unambiguous information exchange between the two makes these systems more reliable, efficient, less error prone, and capable of solving complex tasks. Emotion recognition is a realm of HCI that follows multimodality to achieve accurate and natural results. The prodigious use of affective identification in e-learning, marketing, security, health sciences, etc., has increased demand for high-precision emotion recognition systems. Machine learning (ML) is getting its feet wet to ameliorate the process by tweaking the architectures or wielding high-quality databases (DB). This paper presents a survey of such DBs that are being used to develop multimodal emotion recognition (MER) systems. The survey illustrates the DBs that contain multi-channel data, such as facial expressions, speech, physiological signals, body movements, gestures, and lexical features. Few unimodal DBs are also discussed that work in conjunction with other DBs for affect recognition. Further, VIRI, a new DB of visible and infrared (IR) images of subjects expressing five emotions in an uncontrolled, real-world environment, is presented. A rationale for the superiority of the presented corpus over the existing ones is instituted.

Keywords: emotional corpora; multimodal recognition; emotion in human–computer interaction

Citation: Siddiqui, M.F.H.; Dhakal, P.; Yang, X.; Javaid, A.Y. A Survey on Databases for Multimodal Emotion Recognition and an Introduction to the VIRI (Visible and InfraRed Image) Database. *Multimodal Technol. Interact.* **2022**, *6*, 47. https://doi.org/10.3390/mti6060047

Academic Editor: Cristina Portalés Ricart

Received: 7 April 2022
Accepted: 15 June 2022
Published: 17 June 2022

1. Introduction

HCI attempts to bridge the gap between users and machines. It consists of the interfaces for humans and the new means of interacting with machines. The interaction can be subjected to input, where machines perceive the input in different forms from the user and decipher meaningful information to carry out the required action. Another goal of the interaction is conveying the output from the machine in a usable and human-friendly form. The HCI can be broadly understood as unimodal and multimodal. As names insinuate, unimodal refers to using a single stream of input to the machine. At the same time, multimodal is expounded as a collaboration of more than one input stream emanating from different sensors to provide a more efficient and accurate input. Despite being less ambiguous, unimodal systems always dawdle when compared to multimodal systems. Multimodal systems provide a more natural way of interaction, where the unimpeded choice of modalities results in a more human–human-like interaction. Being redundant increases the odds of the unruffled transmission of information, owing to the presence of simultaneous references conveying the bits of information through different channels for the same task. Multimodal systems are more expressive, efficient, unambiguous, infallible, easier for resolving problems, and have a wider application domain. MER systems, therefore, are known to be more fitting and expressive for affect analysis and recognition.

Unimodal emotion recognition research found its satiation in terms of accuracy and applications when HCI recognized multimodal recognition of emotions as the next trend in this research area. Several modalities, such as facial expressions, speech, physiological signals, and body movements, are now being combined to develop more accurate and physical world-like systems. Humans decipher emotions by combining the signals from different sources and inferring the emotions. A machine detecting the emotions in the same way will bring human and machine interaction (HMI) closer to the desired goal of HCI.

Modalities are being used in different combinations with different DBs to proliferate the accuracy of discerning emotions. One of the most natural and widely used combinations is facial expressions with voice. Works such as [1–27] employed facial expressions in combination with speech prosody to perform bi-modal emotion recognition. A few other works added another modality to solve specific problems, e.g., hand gestures [28,29] to solve the problem of missing data at the fusion stage, and body gestures [30] to perform a feature level fusion and classification using a Bayesian classifier. Similarly, thermal and visible facial images and voice were employed [31,32] for affect recognition using neural networks and hidden Markov models (HMM). An interesting fusion of faces, utterances, and language proposed a framework of MER using undirected topic models and compared the task of emotion recognition to semantic topic analysis in documents [33]. Physiological signals, such as electroencephalogram (EEG), galvanic skin response (GSR), skin temperature (SKT), respiration pattern (RP), electromyogram (EMG), and electrooculography (EOG) were also applied to classify emotion in 13 different categories using feature level fusion [34]. Other physiological signals, such as electrocardiogram (ECG), electro dermal activity (EDA), heart rate (HR), skin conductance level (SCL), and skin conductance response (SCR), were also utilized to detect the emotional state of a person using a series of convolutional neural networks (CNN) and long short-term memory (LSTM) recurrent neural networks (RNN) [35]. The physiological signals were also deployed for detecting and classifying emotions using deep learning (DL) architectures [36,37], and were combined with facial expressions to report emotional recognition accuracy in valence and arousal dimensions [38].

A common aspect in most of the works of emotion recognition is using a DB to train and verify the system. ML algorithms, specifically DL, are employed in various works to apply pattern recognition in emotion detection. Thus, DB forms an integral part of the research. A DB high in quality and size is superior to any level of sophistication incurred in state-of-the-art algorithms. The contributions of this survey paper are the (i) identification, description, and classification of multimodal DBs, (ii) identification of modalities and features involved in emotion detection, and (iii) introduction of a new visible and IR image DB, the VIRI DB, that can be utilized for emotion recognition. The rest of the paper is organized as shown in Figure 1.

Method and Keywords

For the literature survey, several keywords were selected based on the technologies involved and the domains focusing on MER, as listed in Table 1. ML is identified as one of the salient modus operandi for identifying an emotional state. DBs, being imperative to ML, are explored alongside the research works, where they have been put to service. The DBs (DB) assessed in this survey must belong to one of the following categories: multimodal (2+) DB, bi-modal DB, and unimodal DB. The search was limited to the DB housing multiple modalities and single modality DB used in conjunction with other unimodal DB to perform MER. The academic journal DBs employed for this search include IEEE Xplore, ACM Digital Library, Science Direct, Scopus, Google Scholar, and Research Gate. Articles demonstrating relevance to the MER were included in the study. Based on the primary and secondary keywords, the articles were further filtered.

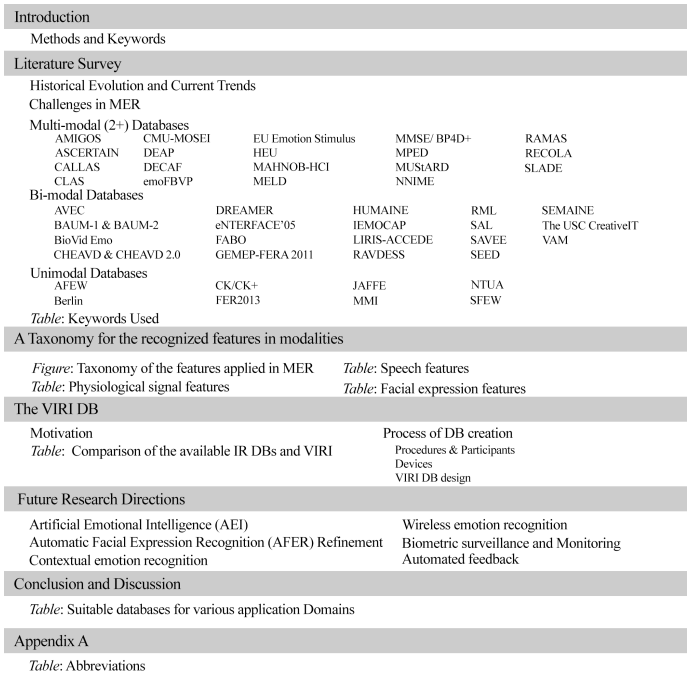


Figure 1. Review paper structure.

Table 1. Keywords used.

Primary Keywords	
Core concept	Keywords
Emotion Recognition	Human computer interaction, human machine interaction, human robot interaction, expressions, expressive, wild, real-time, e-learning, contextual, biometric surveillance, marketing, monitoring, automatic feedback, medicine, intensity, elicitation method, geometric based, FACS coded, AU, continuous, discrete, emonets, sensors, stimuli.
ML	DL, classifiers, ensemble, WEKA classifiers, Naïve Bayes, ANN, CNN, RNN, HMM, DBN, CDBN, DCNN, supervised, unsupervised, wavelet, multiresolution, features, SVM, MLP, kNN, MCC, relational auto-encoders, transfer learning, pre-trained, ResNet, LSTM, RBM, pattern recognition, RF, regression.
DBs	Multimodal, bi-modal, unimodal, corpus, dataset, participants, categories, spontaneous, acted, posed, induced, natural, background, wild, languages, excitation, valence, activation, dominance, arousal, annotated, expectation, power, crowd sourced.
Modalities	Voice, speech, face, gesture, video, audio, body gesture, physiological signals, visible images, IR/thermal images, EEG, GSR, BVP, RP, SKT, EMG, EOG, aligned faces, non-aligned faces, visual features, acoustic features, physical interaction data, visual interaction data, HR, SCL, SCR, EDA, ECG, respiration frequency, pulse, language, acoustic features, lexical features, linguistic cues, acoustic cues, contextual social sciences.
Secondary Keywords	
fusion, feature level, decision level, hybrid fusion, late fusion, early fusion, hierarchical classifier fusion, multiple kernel learning, HOG-TOP, Support Vector Regression, ELM, MDR, WLD, HDFS, social robots, probability, n-fold cross validation	

2. Literature Survey

This section is apportioned into six sub-sections. It provides historical evolution and current trends in the field of emotional recognition along with the challenges. In addition, it also presents the survey of the DBs plied in MER and the works where they have been leveraged. The survey on the DBs is summarized in Table 2 to give a comparison of various attributes toward the end of this section.

2.1. Historical Evolution and Current Trends

Different theories and definitions have been proposed regarding emotion and emotional expressions over the years. Charles Darwin proposed one of the first theories regarding emotional expression in his book *The Expression of the Emotions in Man and Animals*. He proposed that emotional expressions are responses to environmental challenges and are closely related to the survival values [39]. Similar to Darwin, James viewed emotion as an effect of something. According to the theory proposed by James and Langer, the individual experiences an emotion due to a physiological response. However, contrary to James, Cannon believed that emotions are first felt, and then reaction occurs to show certain behavior [40].

The first work on automatic facial emotion recognition (AFER) started in 1978; however, the work was not able to create a high impact in the research community because of the poor performance of the proposed algorithm [41]. Later, during the late 1990s, Paul Ekman and his colleagues studied human facial expressions (FEs) in different cultures. They found both similarities and differences in facial expressions; therefore, they proposed that the facial expressions are governed by the "display rule" in different social contexts [42]. Ekman and researcher Friesen also developed the facial action coding system (FACS) to code facial expressions. Their work became an inspiration to many researchers in analyzing facial expressions using images and videos [43].

Mase was one of the first to use image processing techniques to recognize facial expressions [44]. Along with Mase, all researchers from [45–49] used optical flow (OF) based processing to recognize facial expressions.

Ekman and Friesen, in their work [43], outlined six universal emotions, such as anger, disgust, surprise, fear, happiness, and sadness, as the ones generally shown by humans, regardless of sex, ethnicity, age, and culture. Later, researchers worked on recognizing more acted emotional states, such as joy, pride, shame, defeat, anxiety, sympathy, etc., and discrete universal emotions. However, pointing out the limitations of discrete emotions to accommodate blended feelings in one of its categories, Langer, during the late 1990s, proposed arousal and valence as a new emotion measurement method [50]. The valence value was used to measure pleasure level. The positive value indicated pleasant emotions, such as joy and pleasure, whereas the negative value indicated emotions, such as sad and fear. Similarly, the arousal scale measured the agitation level of emotion. Researchers, over the years, have profoundly used this measurement approach in emotion recognition; however, it has been largely applied to physiological-based emotion recognition. The interested readers can refer to [51–55] for the related work.

The research in AFER received more attention when CK (Cohn-Kanade) data were published in 2000. Most of the early work focused on FACS to code facial expressions where face movements were expressed by either limited or large sets of action units (AUs) [56,57]. Matsugu, during the early 2000s, developed the first facial emotional recognition model [58]. He used the CNN model in his work to find the local differences between neutral and emotional faces. Later with the publication of the BU-3DFE (Binghamton University 3D Facial Expression) dataset in 2006, researchers started extending RGB-based facial emotion recognition to 3D [59].

Besides the previously proposed methods, some researchers also used a curvelet-based feature extraction technique. They represented curvelet by using discontinuities in 2D functions, and those curvelets were used for training the algorithm [60]. Other researchers used 3D faces in their work, where the principal component analysis (PCA) method was used for

feature extraction and support vector machine (SVM) for classification purposes [61]. Extending the work by [60,61], some researchers introduced a new curvelet-based algorithm by combining bilateral two-dimensional PCA feature extraction techniques along with an extreme learning machine (ELM) algorithm [60]. They used the proposed algorithm with different databases such as FERET (facial recognition technology), Faces94, JAFFE (Japanese female facial expressions), Georgia Tech, Sheffield, ORL (Olivetti Research Laboratory), and YALE. Besides the aforementioned techniques and algorithms, where most of the research was focused on FEs and AUs, works on spontaneous facial expression detection, analysis of the complex mental state, detection of human physical and psychological behavior, using both posed and spontaneous datasets, have paved the way to new applications in the field of AFER [41].

Most of the work on AFER during the late 1990s and early 2000s was focused on a geometric-based approach, where methods such as OF, AUs, FEs, curvelet, PCA, etc., were predominant in applications. Different classifiers, such as HMM, k-nearest neighbors (kNN), neural networks, distance-based, rule-based, and Bayesian networks, were used. However, with the introduction of the deep neural network (DNN), researchers introduced a pattern-based approach for AFER. Researchers from [62] used single layer DNN containing convolution layers and deep residual block for the classification of an image into one of six facial emotion classes. Similar to [62], researchers from [63–65] also used DNN for AFER.

Voice-based emotion recognition (VER) has also been done for decades. Traditional methods on VER were based on short-time frame-level feature extraction. Later, researchers started to use utterance-level information extraction; however, lately, deep learning-based feature extraction methods have become more important. Some other traditional approaches in VER also include the extraction of prosodic information, such as pitch, duration, the intensity of utterance, etc. [66]. Apart from these, researchers, during the 2000s, also used such features as acoustic, local invariant features (LIF), affect-salient discriminative features, sparse coefficients, and Fourier parameters combined with machine learning-based classifiers to achieve VER [67–70]. However, researchers have started to use a pattern-based approach in VER in the last few years, where DNN is used for both feature extraction and classification. Researchers from [71] used DNN, which uses a one-second frame of a speech spectrogram to recognize emotions. Similarly, researchers from [72] used convolution and recurrent networks for emotion recognition by learning emotions directly from spectrograms. The interested readers can refer to [8,73] for some more recent work VER using DNN.

Similar to VER, researchers over the years have also used physiological features such as HR, RP, EEG, blood oxygen saturation (BOS), GSR, and SKT for emotion recognition [38,74–76].

Researchers from [77] were the first to apply physiological features to classify emotion, where they attended 38–51 percent accuracy on the classification of four different facial expressions, such as happy, sad, anger, and fear, when four facial EMG signals were given. Lately, researchers from [78] used DNN for emotion recognition, using EEG and peripheral physiological features. Similarly, researchers from [51] used deep CNN on a dataset containing information on GSR and ECG and correlated physiological signals with the data of arousal and valence of the dataset for the detection of the emotion.

The studies of facial expressions, bodily expressions, voice signals, and psychological features have been mainly conducted independently of each other for emotion recognition. However, during the early 2000s, researchers started to explore more techniques, such as fusion of the aforementioned parameters, to enhance the performance of emotional recognition. Researchers in their study [79] found that during one-on-one interaction, 93 percent of our communication may be transferred through paralanguage, such as voice tone, body language, facial expressions, eye movement, etc. In addition to this, researchers [66] also found that the study of only one modality to detect emotion is not enough. For instance, the real emotion behind a person's disagreement hidden by a fake smile cannot be interpreted by just the study of facial expression until other modalities, such as body expressions,

voice, and psychological signals, are also considered. Such findings during the study encouraged researchers to combine more modalities in addition to one or two modalities in their research.

Researchers from [41,58,80], primarily gave classification for emotion recognition using multimodal approaches. They talked about different features and techniques used in their work for emotion recognition. Similarly, researchers from [81,82], during the early 2000s, worked on methods to classify audiovisual inputs into six emotional categories, such as happiness, sadness, fear, anger, surprise, and dislike. They used both audio and visual signals for emotion detection. Researchers from [83] used a fusion of multimodal features, such as facial expression, body gesture, and speech signals to achieve emotion detection. Similarly, researchers from [84] fused 2D and 3D facial features along with speech features for emotion detection. In addition to this, researchers from [85] also successfully fused features such as EEG, pupillary response, and gaze distance to develop a multimode recognition system. However, recently, researchers have started to use DL architectures for emotion recognition. Researchers from [86] applied a deep belief network for multimodal emotion recognition using body, voice, and physiological signal modalities. Similar to [86], researchers from [87] used a convolution deep belief network for multimodal emotion recognition. The interested readers can refer to [8,88] for some more recent work on multimodal emotion recognition.

In summary, though the study of emotion recognition was primarily motivated by evolutionary accounts of emotion, research in MER only started during the late 1990s. Initially, research was focused on geometric-based approaches such as FEs and AUs; however, researchers later started to use pattern-based approaches, such as DNN and CNN. Similarly, though single modalities were used for emotion recognition initially, later researchers started to use the fusion of multiple modalities to enhance the performance of emotion recognition. Moreover, to deal with challenges, such as large pose variations, illumination conditions, and subtle facial behavior, modalities such as 3D, thermal, and IR were also proposed. Hence, over the years, these developments in emotional recognition have paved the way for many real-world applications, such as marketing, medicine, education and research, autonomous driving vehicles, aid for the disabled, HCI, robotics, safety aids, entertainment recommender systems, and automated surveillance.

2.2. Challenges in MER

2.2.1. Limitation in the Background or Emotional Categories of DBs

A large number of unimodal and multi-model DB available today can be used for emotional recognition; however, they pose some drawbacks and limitations. IRIS (Imaging, Robotics and Intelligent System) DB [89] and NIST (National Institute of Standards and Technology) DB [90], which are profoundly used by the research community for emotional recognition, lack data related to all emotional states. Similarly, NVIE (Natural Visible and Infrared Facial Expression) DB [91], on the other hand, lacks all the six emotions triggered simultaneously. Additionally, though KIFE (Kotani Thermal Facial Emotion) DB [92] has all of the emotion states recorded, it has been performed in a controlled laboratory environment. There is also the scarcity of the DB for research-related intensity measurement. Therefore, the aforementioned limitations make the implementation of emotional recognition in a wild environment difficult. Moreover, the emotional recognition model built with such data may not be robust and accurate to detect all the emotional states.

2.2.2. Real-Time Implementation

Real-time implementation of emotion recognition has always been challenging due to time resolution, low-level emotion recognition, face and angle variation, illumination variations, and non-alignment of the faces. Therefore, such drawbacks prevent potential implementation of emotional recognition in real-time applications, such as patient monitoring, surveillance, security, and biometrics. Moreover, most of the database available today lacks lower intensity level features and are concentrated on emotions shown at the peak

level of intensity. Therefore, measuring the low-level fame intensity has been challenging, as the available database lacks such discriminating features.

2.2.3. Dataset Categorization

Datasets are commonly divided into two major categories: spontaneous and posed. Posed datasets are popular for capturing extreme emotions, whereas spontaneous datasets capture natural human behaviors. Datasets such as JAFFE, CK, Bosphorus, BU-3DFE, IRIS, NIST, NVIE, KTFE, etc., contain posed expressions. In contrast, datasets such as CK+, MMI (Michel Valstar and Ioannis Patras), etc., contain spontaneous expressions. Moreover, these datasets are further categorized into RGB, thermal, and 3D. It is challenging to choose a particular database for a specific application with these many variations and categorizations.

2.2.4. Emotional Annotation

Emotional label annotation is challenging, as it depends on the annotator's perspective. Self-assessment by the person involved in the conversation is possibly the best way to annotate utterances; however, it is not possible because it will impact the conversation flow. In some datasets [93], annotators are people uninvolved in the script and conversation, who do not know the proper emotion behind the utterances. In such datasets, data and values related to emotional intensity and features during a conversation might not reflect the correct notion or emotional state behind it.

2.3. Multimodal (2+) Databases

2.3.1. AMIGOS

A dataset for the multimodal research of affect, personality traits and mood on individuals and groups (AMIGOS) is a multimodal DB consisting of neural and peripheral physiological signals (EEG, GSR and ECG), video and depth information [94]. The authors accumulated data in a set of two experiments in two social settings, one in the individual context and the other in the group context, using emotional movie excerpts (of short duration, <250 s and of long duration, >14 min) as the elicitation method. The data were recorded in individual settings and group settings to observe and analyze the impact of social context on a person's emotional state. It was observed that for lower arousal clips, group settings reported lower levels of arousal as well. In addition, for higher arousal clips, group settings reported higher arousal, and individual settings reported lower valence levels. The data were recorded using wearable sensors for the neurophysiological signals. Emotiv EPOC Neuroheadset4 (14 channel, 128 Hz, 14-bit resolution) [95] was used to record EEG signal, and the Shimmer 2R5 platform extended with an ECG module board (256 Hz, 12-bit resolution) [96] was used for recording the ECG signals. Videos of the frontal face were recorded in high-definition using a JVC GY-HM150E camera, while Microsoft's Kinect V16 was used to document the RGB and full-body depth videos. The author tested the dataset for classification using the Gaussian naive Bayes classifier in different settings. The result was reported in the form of mean F1-scores (EEG only: 0.564 for valence (V), 0.577 for arousal (A), GSR only: 0.528 (V),0.541 (A), ECG only: 0.545 (V), 0.551 (A), Fusion (EEG+GSR+ECG): 0.560 (V), 0.564 (A)).

2.3.2. ASCERTAIN

A multimodal databaASe for impliCit pERsonaliTy and Affect recognitIoN (ASCERTAIN) comprise physiological signals data (EEG, ECG and GSR) together with facial activity data [97]. The authors claim it to be the first one to provide the potential for emotion recognition and personality traits. Another distinctive feature of the corpus is wearable off-the-shelf sensors for capturing the physiological signals and a web camera for recording facial landmark trajectories (EMO) that makes it viable for use in commercial applications. The authors tested the DB for affective state recognition using SVM and naive Bayes classifiers, and the results were reported in the form of mean F1 scores. The best

score was achieved for a late fusion of all four modalities (valence: SVM = 0.69, NB = 0.71; arousal: SVM = 0.64, NB = 0.67).

2.3.3. CALLAS Expressivity Corpus

Conveying affectiveness in leading-edge living adaptive systems (CALLAS) DB is a multimodal corpus of cross-cultural affective behavior aimed at discerning patterns comprising the non-verbal behavior across cultures [98,99], namely, German, Italian and Greek cultures. It was created within the European Integrated Project CALLAS and comprises the modalities of facial expressions, speech, and body gestures. The prime modality of the DB is hand gesture expressivity, and facial expressions and voice utterances were recorded to support the claims of the principal modality. The corpus was constructed using multitudinous human behavior capture methods, such as video, Wii mote, and data gloves. An outcome of the study was observing a pattern that German gestures were more unhurried than Italian or Greeks and were also executed with less activity.

CALLAS creators claimed that the corpus was not meant to study cultural diversity in detail, and they only used a sub-corpus of German contributors. An estimated 5 h of interaction was recorded using two cameras (for face and the whole body), and a microphone that included reading and device-changing time. This DB was used for MER using voice, face and gestures to report results for an individual unimodal system (42–51% accuracy) and multimodal system (≈55% accuracy) [28].

2.3.4. CLAS

A database for **c**ognitive **l**oad, **a**ffect and **s**tress recognition (CLAS) is a multimodal DB that was developed for the automatic recognition of some states of mind, particularly negative emotions, high cognitive effort and mental strain [100]. The experiments elicited emotions that included perceptive tasks (2) and interactive tasks (4). The role of the perceptive tasks was to invoke emotions using images and audio–visual stimuli. The interactive tasks were carried out to elicit various cognitive efforts. This was done by asking the participants to solve math problems, logic problems and the Stroop test. The participants were asked to complete a questionnaire after each experiment. The three physiological signals, PPG, ECG and EDA, were recorded by Shimmer3 ECG Unit and Shimmer3 GSR+ Unit [101,102]. The Shimmer3 GSR Unit captured the 3D accelerometer data. The analysis of the success rate in the interactive tasks, the responses in the questionnaires, and the physiological signal reading facilitated the evaluation of specific states of mind. The authors reported some baseline results in the form of valence and arousal values using SVM-based detectors. The classifier was fed with two combinations, and the best result was reported for a perceptive task using picture stimuli. It reported the best arousal value for PPG+GSR (77.6%) and best valence for ECG+GSR (74.5%).

2.3.5. CMU-MOSEI

The Carnegie Mellon University multimodal opinion sentiment and emotion intensity (CMU-MOSEI) is a multimodal dataset for sentiment analysis and AER [103]. The bulk of the corpus is from reviews (16.2%), debates (2.9%), and consulting (1.8%). The annotation of the DB was carried out by three crowd-sourced judges (from the Amazon mechanical Turk platform).

2.3.6. DEAP

Database for emotion analysis using physiological signals (DEAP) is a multimodal DB of 32 participants designed to assist the analysis of the spontaneous human affective state [104]. DEAP was used in a DL architecture for MER using DBN and CDBN that achieved an accuracy of 79.5% [86]. The DB was also employed for a multimodal feature-level fusion framework using physiological signals for emotion classification of 13 emotions [34]. DEAP was also utilized for testing emotion assessment by the proposed multiple-fusion-layer based ensemble classifier for stacked autoencoders (MESAE) [36].

2.3.7. DECAF

A multimodal dataset for DECoding user physiological responses to AFfective multimedia content (DECAF) comprises brain signals, near-infrared facial videos and peripheral physiological signals [105]. The corpus compilation was done to evaluate emotion recognition of the magnetoencephalogram (MEG) against EEG sensing and between movie and music clips. A MEG sensor was employed to record the brain signals (a distinct characteristic of MEG is that it requires little physical contact with the user's scalp). The physiological signals that were recorded are tEMG, hEOG, and ECG. A device, ELEKTA Neuromag, was used to synchronize MEG and physiological signals, and the sound of the stimulus videos was recorded and used to synchronize the user's facial NIR videos. The author's observation for the comparison of affect encoding by MEG and EEG stated a very comparable results, yet the higher spatial resolution of MEG allows for further substantial analysis of affect that may enable higher recognition performance. The comparison of the stimuli for emotion recognition showed better F1-scores using MEG features for movie clips. This manifested movie clips as a better means than music to instigate emotions.

2.3.8. emoFBVP

The emotion face body gesture voice and physiological signals (emoFBVP) DB comprises audio and visual sequences of emotions along with physiological signals of the actors [86]. The DB includes facial expressions, body gestures, speech, and physiological signals recorded simultaneously using the Microsoft Kinect, the Zephyr BioHarness, and wrist accelerometers. The corpus is well suited for different combinations of uni-, bi- or multimodal systems. The corpus has been used for the unsupervised classification of 23 emotions using DBN and CDBN, achieving the highest accuracy of 83.18% [86] compared to the use of other DBs, such as CK, DEAP, and MAHNOB-HCI [85]. An extension of this work utilized emoFBVP for examining the transfer of emotion features between two artificial neural networks (ANN), where a comparable accuracy with less training time was reported for emoFBVP.

2.3.9. EU Emotion Stimulus

Constructed as a part of the "ASC (Autism Spectrum Condition) Inclusion project" within the European Community's Seventh Framework Program, the EU Emotion Stimulus Dataset is a multimodal DB of facial expressions, speech, body gestures and contextual social scenes [106]. It is a collection of 20 (+a neutral) dynamic mental and emotional states recorded from an ethnically varied group of young and adult actors. This DB was created to fulfill the need for an emotion stimulus set required to develop an online socio-emotional training tool for children with a diagnosis of ASC. This DB was used for an enhanced emotion recognition system using CNN and restricted Boltzmann machines (RBM), where shallow RBM was employed for the arousal/valence classification of emotions using a 5-fold cross validation [10].

2.3.10. HEU

The Harbin Engineering University (HEU) [107] DB comprises one of the most extensive corpora for MER in the uncontrolled wild environment. The DB is in the form of video clips and consists of HEU-part1 (downloaded from online sources, such as Giphy, Google, and Tumblr) and HEU-part2 (taken manually from TV series, movies, and live shows). The creators of the DB evaluated the DB on conventional machine learning and deep learning architectures. They reported an increase in the accuracy of MER over a single modality (facial expression) by 2.19% and 4.01%, respectively.

2.3.11. MAHNOB-HCI

The multimodal analysis of human nonverbal behavior: human–computer interfaces (MAHNOB-HCI) [85] DB exhibits data for six modalities (32 channel EEG, physiological signals, facial expressions, audio, eye gaze, and body movements). The DB was devel-

oped to support emotion recognition using a single modality and through the fusion of modalities to predict a correct emotion. Two experiments were performed, where the first one involved watching 20 emotional videos and self-reporting their feelings. The second experiment involved viewing 14 excerpts and 28 images without and then with the wrong tags. The user input regarding agreement or disagreement with the tags and the video and body movement were segmented and recorded. This DB was used for multimodal recognition using deep belief networks (DBN) and convolutional deep belief networks (CDBN) in unsupervised classification, with reported accuracy being 58.5% [86]. This DB was also used for reporting results on unimodal as well as a fusion of multiple modalities (facial expressions and EEG) in [108], and it was observed that fusion at the decision level produced better results than feature-level fusion. The DB is publicly available via a web-based system for MER models.

2.3.12. MELD

The multimodal EmotionLines dataset (MELD) [109] DB is a multimodal emotional corpus consisting of conversational videos from the American sitcom, "Friends". This DB has evolved from its predecessor EmotionLines [110], which consisted only of conversations. As an extension and enhanced version, MELD contains the audio, video, and textual data of each utterance. The DB has been used extensively for emotion recognition in conversation (ERC). A work mentions using MELD for testing Sentic GAT, a context- and sentiment-aware framework [111]. The DB has been employed for ECR using the long short-term memory (LSTM) neural network by tapping the emotional shift detection (ESD) as a guiding principle for MER.

2.3.13. MMSE/BP4D+

The multimodal spontaneous emotion database (MMSE) is a multimodal DB comprising a diverse set of modalities [112]. Participants were asked to conduct 10 activities, under the supervision of a professional actor, in a natural transition from positive toward negative emotion. The 3D geometric sequences and the simultaneous 2D videos models were captured using the Di3D dynamic imaging system (one 3D sensor + one 2D RGB camera + two monochrome camers). For capturing the thermal videos, a FLIR3 A655sc Longwave infrared camera was employed (resolution = 640 $\times$ 480) and a Biopac MP150 data acquisition system was used to accumulate the physiological signals data. After completing each task, every participant self-reported the annotation for each activity they carried out in terms of the feeling he/she experienced. They reported the emotions as well as the intensities, ranging from very slightly to extremely. The data were also annotated by expert FACS coders for all the participants. The five experts encoded a total of 34 facial action units (AU), while two experts encoded intensity on AU for a subset of the DB. A mention of MMSE comes as BP4D+ DB in research for AER by facial expressions based on de-expression residue learning (DeRL) [113]. Only the 2D texture component of the database was used for pre-training the model. The term de-expression was coined for the process of creating a neutral face image for any image that is provided as an input by the generative model trained by cGAN. The method then learns the residue (the deposition) that remains in the generative model. An accuracy of 81.39% was obtained for a four expression classification using MMSE DB for training and testing. Another previously mentioned method also utilized MMSE for the AER using audio–visual data obtained from a mobile phone [114].

2.3.14. MPED

The multimodal physiological emotion database (MPED) [115] is a multimodal corpus consisting of simultaneous recordings of four physiological signals (EEG, GSR, respiration, and ECG). The DB was developed by eliciting subjects' emotions using 28 emotional videos that humans manually annotated. The DB has been used in a novel graph-embedded convolutional neural network (GECNN) method for emotion recognition using the EEG

signals [116]. The authors proposed a method to convert the discrete EEG signals into continuous images and fusing the global and local functional features for an image-based EEG emotion recognition.

2.3.15. MUStARD

The multimodal sarcasm detection (MUStARD) [117] DB is a corpus for the detection of sarcasm, "a thinly veiled attempt to disguise feelings of anger, fear, or hurt " [118]. The DB consists of audiovisual cues (in the form of video clips from American sitcoms "Friends", "The Golden Girls" and "The Big Bang Theory", and the Scottish TV show "Sarcasmaholics Anonymous") accompanied by the textual context of each utterance. A total of 690 videos were selected from the initial pool of 6365 annotated videos. This initial pool included 400 videos from the MELD [109,110] DB for the non-sarcastic category. The DB was annotated by two judges (a third judge for breaking the tie), who used a web interface to classify the video, accompanied by the transcript, into sarcastic and non-sarcastic categories. The authors conducted experiments using the MUStARD DB and reported a reduction in error rate by 12.9% when using the multimodal variants of the DB over the unimodal counterparts. The DB has been used for multimodal sentiment classification and MER using an external knowledge-enhanced multi-task representation learning network (KAMT) [119]. The DB has also been employed to analyze a multimodal learning system, multimodal learning using optimal transport (MuLOT) [120], to detect sarcasm and humor.

2.3.16. NNIME

The National Tsing Hua University-National Taiwan University of Arts Chinese Interactive Emotion (NNIME) [121] is a multimodal DB in the Chinese Language. The DB has been developed as a collaborative work between the National Tsing Hua University and the National Taiwan University of Arts. The DB is a simultaneous recording of audio, video, and ECG signals that were carried out as spoken dyadic interactions (as pairs performing a spontaneous and short scene of about 3 min) between 44 subjects under the supervision of a professional director. The DB consists of both discrete and continuous emotion annotations carried out by 49 annotators consisting of students and professors. The DB has been used for emotion recognition using only speech by considering the nonverbal vocalization in conversations depicting emotions [122]. They used only the audio portion of the DB and used an LSTM to acquire the shifts in the dialogue of the speaker's emotion from a sequence of segmented speech signals.

2.3.17. RAMAS

The Russian acted multimodal affective set (RAMAS) [123] DB is a multimodal corpus for MER in the Russian language. Neurodata Lab LLC created the DB, where 10 semi-professional actors played dyadic scenarios to depict each of the basic emotional categories. Twenty-one annotators annotated the records using the Elan [124] tool from Max Planck Institute for Psycholinguistics (the Netherlands). The RAMAS DB has been used for testing a bi-modal emotion recognition system that avails the acoustic and linguistic information of speech signals [125].

2.3.18. RECOLA

An affective and collaborative interaction dataset, remote collaborative and affective interactions (RECOLA), was recorded at the Department of Psychology at Universit'e de Fribourg-Universit (Freiburg, Switzerland) synchronously for multiple input channels of speech, face, ECG, and EDA [126]. The participants were paired as 23 teams and asked to complete a task requiring collaboration while being video-recorded. The participants self-assessed the tasks for both social and emotional behavior, which were then annotated by six annotators who measured the corpus on valence and arousal for emotional behavior and five dimensions for social behavior. While emotional behavior showed an agreement

between the participants and the annotator's observation, the social behavior accord was not sound.

RECOLA facial expressions, audio features, and physiological signals were used to predict emotions in valence and arousal dimensions, with the result being computed as a concordance correlation coefficient (CCC) and reported to be 0.683 (arousal) and 0.642 (valence). The CNN and deep residual networks (DRN) were used to predict the emotions, using visual and speech modalities only [8]. The DB was tested for the fused outputs, and accuracies of 0.612 (valence) and 0.714 (arousal) were observed. In another work for affective observation using physiological signals, this DB recorded accuracies of 0.430 (arousal) and 0.407 (valence) [35]. In a single-subject multimodal regression model (SSMRM), RECOLA was used for face, speech, ECG, and EDA signals [127]. The results were obtained for all modality combinations, and the best accuracy for arousal was obtained using only audio. In the case of valence, the combination of facial expressions and speech produced the best output.

2.3.19. SLADE

A multimodal DB of physiological signals, the stress level and emotional state assessment database comprises 16 channel EEG signals, 3 channel ECG signals, GSR signals and SKT measurements to assist AER [128]. The main aim of this DB development was to facilitate the detection of stress levels and concurrently the recognition of emotional states through synchronized physiological signals. The EEG signals were procured using a 16-channel Ultra Cortex 3D printed headset and OpenBCI (open-source platform for brain computer interfacing) board integrated with a Neurosoft elastic headset. The Vernier ECG sensor acquired the ECG signals. The authors captured the remaining data using a custom module developed by them at the Sensor Network Laboratory (Technical University of Varna). The data were annotated by each participant collected as self-assessment during the experiments. The authors performed a baseline study for the stress events in context-specific and application-specific scenarios and reported results for automatic detection in HALV (high arousal/low valence) events. An accuracy of 92.9% for a binary valence detector, 100% for a binary arousal detector, and 100% for the HALV detector was achieved for the combination of all four physiological signals.

2.4. Bi-Modal Databases

2.4.1. AVEC

Developed for use in emotion and depression recognition challenges, the audio visual emotion challenge (AVEC) DB is a bi-modal corpus of facial expressions and speech [129,130]. Subjects performed HCI tasks while they were being video-recorded. Each subject was recorded 1–4 times, two weeks apart. The HCI tasks included scenarios, such as sustained vowel phonation, counting from 1 to 10, reading aloud, telling a story from the subject's past, and best present ever. For the 2014 AVEC challenge, two tasks in German, Northwind, and Freeform, were selected. The Northwind included reading the excerpt of the fable "*Die Sonne und der Wind*" (the north wind and the sun), while Freeform required participants to respond to several questions, such as discussing a sad childhood memory, etc. AVEC DB was used for a deep bi-directional LSTM-RNN (DBLSTM-RNN) based uni- and multimodal emotion recognition system that reported the valence, arousal, and dominance for Freeform, Northwind, and Freeform–Northwind development sets. The Northwind task achieved the best average accuracy of 0.630 [20]. In a hierarchical fusion strategy for MER, AVEC DB was used for the hierarchical level fusion of acoustic and lexical features using LSTM models, where an accuracy of 69.9% was achieved [131].

2.4.2. BAUM-1 and BAUM-2

The Bahcesehir University multilingual (BAUM-1) DB is a spontaneous audio–visual DB that comprises the subject's affective and mental states in Turkish [132]. The mental states captured in the DB include being unsure, thinking, concentrating, and being bothered.

The subjects were asked to observe formulated image sequences and short clips to kindle a set of emotional and mental states and report sentiments and implications extemporaneously, in their words. The DB poses challenges for use in emotion recognition; however, it serves as an apt corpus for affective systems that seem to recognize emotions in real-world scenarios. BAUM-1 DB was used for testing a hybrid deep model consisting of CNN, 3D-CNN, and DBN for an audio–visual affective system for emotion classification and reported accuracies of 42.26% and 50.11% for audio and visual data, respectively [23]. However, the fusion of the two modalities provided an accuracy of 54.57%. BAUM-2 extends BAUM-1 by the addition of (i) subjects of different races and ages (5–73 years) in various illumination conditions and head poses, (ii) English language using clips from genres, such as crime, thriller, and comedy, (iii) neutral emotion, (iv) profuse annotation conducted by 5 annotators, and (v) gender, age, head-pose and intensity of emotion. BAUM-2i, another extension of BAUM-2, was created by extracting frames from each clip displaying subject emotions at their peak in real-life scenarios. The rich annotation in these BAUM-2 and 2i helps in carrying out affect recognition with varying intensities.

2.4.3. BioVid Emo

The BioVid Emo [133] DB is a multimodal DB comprising simultaneously recorded physiological and video signals. Developed in a collaboration of the Neuro-Information Technology group of the University of Magdeburg and the Emotion Lab of the University of Ulm, this DB was created by eliciting the subjects' emotions when they were asked to watch 15 standardized film clips. The subjects had to choose a clip that elicited the maximum asked emotion and rated the clips based on valence (unpleasant to pleasant) and arousal (calm to excited/activated). The DB has been used to analyze a recurrent model for predicting a person's emotional intelligence by forecasting the pattern of age, gender, occupation, marital status, and education [134].

2.4.4. CHEAVD & CHEAVD 2.0

The CASIA (Chinese Academy of Sciences Institute of Automation) natural emotional audio-visual DB (CHEAVD) [135] is a multimodal corpus in the Chinese language consisting of audio and video modalities. The DB is distinctive, as it consists of 26 uncommon emotional states (neutral, angry, happy, sad, worried, anxious, disgust, surprise, blamed, sarcastic, aggrieved, curious, fearful, embarrassing, nervous, confused, proud, helpless, hesitant, contemptuous, frustrated, serious, anticipated, shy and guilty) with multiple emotion labels to reduce ambiguity in emotion classification. There is an advanced version of the DB, CHEAVD 2.0 [136] that was introduced as a benchmark for the Multimodal Emotion Recognition Challenge (MEC) 2017. Both the versions of the DB are highly uneven, with most of the excerpts depicting neutral emotion. The DBs have been used for emotion recognition using speech by the fusion of features extracted from English and Chinese languages [137] using local Attention RNN (LA + RNN) with a local attention mechanism. The CHEAVD 2.0 was also utilized for testing a cross-culture MER [138] for an adversarial learning framework for generalizing the scenarios across different cultures.

2.4.5. DREAMER

The DREAMER dataset (database for emotion recognition through EEG and ECG signals) is a bi-modal DB comprising 14-channel EEG and 2-channel ECG signals [139]. The EEG signal was recorded using an Emotiv EPOC system (Sampling Rate = 128 Hz) [140,141], a wireless headset containing contact sensors, and the ECG signal was recorded using a SHIMMER wireless sensor (recorded at 256 Hz) [142]. The authors established a standard of their corpus by performing classifications based on supervised methods (SVMs) for single modalities and fusing these modalities. The best result of Valence = 0.6249 for EEG only, Dominance = 0.6184 for EEG only, and Arousal = 0.6232 for the fusion of EEG and ECG was obtained.

2.4.6. eNTERFACE'05

eNTERFACE'05 is a corpus of facial expressions and speech that was developed for testing and assessing single and bi-modality emotion recognition systems [143]. In the DB, ≈31% of people wore glasses, and 17% had a beard. This DB was employed for testing an MER using facial and audio features, where two fusion schemes were assessed—weighted sum rule and weighted product rule—achieving an accuracy of 75.7% and 77.2%, respectively [11]. Another bi-modal system employing audio and video used eNTERFACE'05 to test the fusion of classifier predictions for emotion detection, achieving an overall accuracy of 99.52% [27]. In another work, multi-directional regression (MDR) audio features and Ridgelet transform-based facial image features were combined using the Bayesian sum rule to achieve an accuracy of 83.06% [14]. Another work used this DB for emotion recognition using the Bayesian sum rule for fusion, MDR features for audio, and Weber local descriptor (WLD) for facial images, producing an accuracy of 83.10% [15]. Another bi-modal emotion recognition system that made use of the sparse kernel reduced-rank regression (SKRRR), a linear extension of traditional reduced rank regression (RRR), reported an accuracy of 87.02% for fusion using a support vector machine (SVM) classifier [17]. For multimodal emotion detection, score-level fusion was performed using deep spatio-temporal features [21]. A 3D-CNN (C3D) cascaded with DBN produced an accuracy of 89.39% for fusion using this DB. At the same time, a shared learning approach by enhanced sparse local discriminative canonical correlation analysis (CCA) made use of this DB to achieve an accuracy of 80.1% for six emotions [22]. Another hybrid deep model composed of CNN, C3D, and DBN used eNTERFACE'05 and reported 85.97% accuracy.

2.4.7. FABO

The face and body gesture (FABO) [144] DB is a bi-modal DB consisting of facial expressions and (upper) body gestures. Developed at the University of Technology, Sydney (UTS) in 2005, this DB was created using two cameras that simultaneously captured body gestures and facial expressions. The participants were asked to enact specific emotions elicited by the narration of an emotional situation. The subjects in the corpus were from varied ethnicities, such as European, Middle Eastern, Latin American, Asian, and Australian. The FOBO DB has been used for MER using a two-layer CNN-transformer deep learning technique [145] and reported an accuracy of over 90% for FOBO DB and CK+ DB [146].

2.4.8. GEMEP-FERA 2011

The Geneva multimodal emotion portrayal-facial expression recognition and analysis (GEMEP-FERA) dataset is a bi-modal DB of facial expression and speech [147,148]. About 1260 clips of this DB were used in a study to assess the inter-judge accuracy of DB in terms of reliability and recognition and to augment the plausibility and the portrayal of intensity for a reliable categorization. The GEMEP-FERA corpus was used in facial expression recognition using a novel feature descriptor called the histogram of oriented gradients from three orthogonal planes (HOG-TOP) [16]. The study proposed a new geometric feature derived from the wrap transformation of facial landmarks to detect facial configuration changes. A feature-level fusion for video-based expression recognition resulted in an accuracy of 54.2%.

2.4.9. HUMAINE

Developed to fulfill a need for collecting and annotating affective content, the human–machine interaction network on emotion (HUMAINE) DB is a naturalistic corpus comprising facial images and speech. It contains a rich set of structured labels showing core signs in speech, facial features, and gesture descriptors [149]. HUMAINE clips were annotated for natural and induced stimuli at two levels. At the first level, the global level, the emotional episode is labeled with a global label. Then at the second level, frame level, it was labeled to be time aligned. This facilitated capturing the perceived flow of emotion. The DB has also been formulated at two levels. The first level extracts a section of the recording as a

clip, and the second level involves selecting 50 of those clips. The clips cover wide-ranging material, such as different contexts, emotional space, intensities, and cues from face, voice, gestures, etc. This DB was used in speaker-dependent scenarios, where both feature and decision level fusions were performed for classifying seven emotions [25].

2.4.10. IEMOCAP

Developed at the Speech Analysis and Interpretation Laboratory (SAIL), University of Southern California (USC), the interactive emotional dyadic motion capture (IEMOCAP) dataset is a bi-modal corpus containing facial expressions and hand movements [93]. The authors claim that the DB includes direct and expounding motion capture information, facilitating an in-depth description of gestures that were not available anywhere at the time. For DB creation, the subject pairs were asked to perform three emotional scripts in hypothetical scenarios with markers on their face, head, and hands. A total of 53 markers were attached to capture the subject's body movement, hand gestures, and head movements. The corpus presents detailed information on the motion capture. This DB was used for training a system of audio–visual emotion recognition using DL, where four emotions were classified using four different DBN models, and the best accuracy of 65.89% was achieved using FS-DBN2 (a two-layer DBN with feature selection prior to the training) model [13]. An energy-based variant of RBM, replicated softmax model (RSM), made use of this DB to recognize emotions using a decision-level fusion of face, speech, and language, and achieved an accuracy of 68.92% [33]. IEMOCAP was also used for emotion recognition in spoken dialogues, using a fusion of acoustic and lexical features through a hierarchical fusion strategy where modalities were combined at different levels to achieve 51.3% accuracy (using LSTM) [131].

2.4.11. LIRIS-ACCEDE

The **L**aboratory of **I**nfo**r**matic in **i**mage and **s**ystems information-**a**nnotated **c**reative **c**ommons **e**motional **d**atabas**E** (LIRIS-ACCEDE) is a bi-modal DB consisting of audio–visual modalities [150,151]. The database is rich in diversity and has been annotated (in terms of valence and arousal) using crowd-sourcing through a pair-wise video comparison protocol carried out by 1517 annotators from 89 countries. The DB has been sorted independently around the valence and arousal axis due to two experiments carried out on crowd sourcing.

2.4.12. RAVDESS

The RAVDESS (the Ryerson audio–visual database of emotional speech and song) is a bi-modal DB comprising facial data and vocal data in the form of speech and song [152]. The corpus contains the two modalities in three combinations: facial expressions and voice, only facial expressions, and only voice. A total of 247 participants validated the corpus, and each rated a subset of the whole corpus based on accuracy, intensity, and authenticity. Another group of 72 participants was involved in the reliability task for testing–retesting the data. The database has been made publicly available under a creative commons non-commercial license. A real-time emotion recognition system made use of the RAVDESS dataset's speech component to ascertain emotion in a live recorded speech by inspecting the tonal properties of the utterance [153]. The authors used gradient boosting, SVM, and KNN for four basic emotions (sad, happy, neutral and angry), reporting an accuracy of 61% for gradient boosting, 81% for SVM and 63% for KNN. Another mention of the database comes in audio-linguistics embeddings for capturing the linguistic and acoustic content of a sentence [154]. The authors made two contributions, firstly by progressing to sentence-level embedding starting from the phoneme level through character and then word-level representation, then secondly evaluating these embeddings for speech recognition and AER using the RAVDESS dataset. Another work that fused audio and visual streams to detect emotions from the data captured through a mobile phone made use of the speech component of the RAVDESS [114]. The speech and songs of the RAVDESS were used to train the CNN, while the BP4D+ (MMSE) DB [18] was used to train for images. They

reported accuracies of 99.22% for images, 66.41% for speech and 96.09% for the fusion of images and audio for the two emotional categories.

2.4.13. RML

The Ryerson Multimedia Lab (RML) DB includes facial expressions and speech collected at Ryerson University [155]. The participants were provided with 10 sentences for each emotion and asked to exhibit emotions naturally by recalling any personal emotion-related event. RML DB was used with a 2-stage network containing two deep CNNs (DCNN) and the fusion of these DCNN outputs, resulting in an accuracy of 72.18% with speech and face [18]. The authors also reported an accuracy of 80.36% for audio–visual emotion recognition while employing a hybrid deep model architecture [23]. Another use of the corpus was made in real-time bi-modal emotion detection in a mobile context, where seven emotions were classified, and the results were reported in terms of precision (90.8), recall (90.7), and F1-measure (90.7) for a feature level fusion [156].

2.4.14. SAL

The SAL (sensitive artificial listener) is an audio–visual DB comprising the recordings of the conversation between a machine and a human [149,157]. SAL also has an interface for portraying the machine conversation and is controlled by an operator with four personalities (Poppy, Obadiah, Spike, and Prudence) that users can choose from. These personalities have their emotional trait imbibed in them, e.g., "Poppy is happy, Obadiah is gloomy, Spike is angry, and Prudence is pragmatic". While conversing with a character, the users are drawn to the character's state of mind by exhibiting their emotional state and non-verbal expressions. The data contain varied emotional content in 491 recordings, but with low intensities. SAL DB was also translated into Hebrew and Greek at Tel Aviv University and the National Technical University of Athens, respectively.

2.4.15. SAVEE

Developed in the Center for Vision, Speech and Signal Processing (CVSSP) 3D vision lab, University of Surrey, the Surrey audio–visual expressed emotion (SAVEE) dataset was developed as a requisite for the creation of an emotion recognition system [158]. The subject's reaction was recorded using a 3dMD dynamic face capture system and Beyer dynamic microphone signals. A total of 10 evaluators assessed the DB for accuracy and quality check. SAVEE DB was used for exploring the sources of temporal variation in human audio–visual behavioral data by introducing temporal segmentation and time-series analysis techniques [19]. In a bi-modal fusion of linguistic and acoustic cues in speech, SAVEE was used for affect recognition at the language level using both ML and valence assessment of the words for the classification of 7 emotions [156]. In an affective human–robot interaction, the real-time fusion of facial expressions and speech from SAVEE using 3 DBNs (two for classifying and the third for fusing the o/p of the first two) resulted in an accuracy of 96.2%.

2.4.16. SEED

The SEED (SJTU (Shanghai Jiao Tong University) emotion EEG) dataset constitutes a corpus to ascertain the emotional state of a person by the EEG signals [159–161]. There exist more versions of the SEED DB, such as SEED-IV [162], SEED-V [163], and SEED-VIG [164], that was developed during various research works. The participants themselves annotated the databases during the course of the experiments. The participants were then asked to watch video clips and report their feelings in a sequence of 5 s for the hint about the clip, 45 s for the feedback, and 15 s for the rest between each trial. The dataset was cumulated in the form of EEG signals recorded with an ESI NeuroScan system (Sampling rate = 1000 Hz) from a 62-channel electrode cap. For SEED-IV, SMI eye-tracking glasses were used to record eye movements in the sessions.

SEED-VIG of the SEED dataset comprised similar modalities, but were collected in a distinct style for vigilance estimation [164]. In the form of a 4-lane highway scene, a VR-based driving system was incorporated to collect the EEG and EOG signal, where participants were asked to drive in simulated environment being inattentive. The driving system was designed to induce fatigue in the subjects easily, and the data were collected using the Neuroscan system (sampling rate = 1000 Hz) and SMI eye tracking glasses.

A multimodal emotion recognition framework employing deep canonical correlation analysis made use of SEED, SEED-IV and SEED-V datasets to test their model [160]. They transformed each modality separately and then coordinated different input types into a hyperspace by using specified CCA. They reported an accuracy of 94.58% on the SEED dataset, 87.45% on the SEED-IV dataset, 83.08% on the SEED-V dataset, 88.99%, 90.57%, and 90.67% for three binary classification problem on the DREAMER dataset [139] and on the DEAP dataset [104] for two binary classification problems (accuracy = 84.33% and 85.62%) and for a four-category classification problem with an accuracy of 88.51%.

2.4.17. SEMAINE

The sustained emotionally colored machine–human interaction using nonverbal expression (SEMAINE) DB was created as a recording of the emotion-bearing conversation between a person and an operator simulating a SAL [157,165]. It is an audio–visual corpus built in various configurations, such as Semiautomatic SAL and Solid SAL. The DB was annotated by 6–8 raters and contains supplementary information, such as facial action coding system (FACS) annotation, nods, shakes, etc., for a few recordings.

2.4.18. The USC CreativeIT

The University of Southern California (USC) CreativeIT DB is a bi-modal corpus of speech and body movements created through the collaboration of researchers and theater experts [166,167]. The DB contains the data of 16 (8 female, 8 male) actors conversing in pairs for ≈3 min, eliciting their emotions that were recorded for full-body motion capture, audio, and video. A total of 50 such interactions produced 90 data samples (10 missing audio recordings). The DB was annotated with valence, activation and domination on a scale of 1–5 by at least 3 raters for each actor. The labels are available for activation and valence only and are based on the average of the ratings. This DB was used in the study of affect recognition by integrating cross-lingual emotion information through the fusion of multiple emotion perspectives and reported accuracies of 0.507 and 0.577 for valence and activation dimensions, respectively [167].

2.4.19. VAM

The *Vera am Mittag* (VAM) German audio–visual spontaneous speech DB was created by the emotion research group at the Institut für Nachrichtentechnik of the Universität Karlsruhe (TH), Germany [168]. The DB is a 12 h recording of a German talk show, *Vera am Mittag*, where the discussion between the guests provides a spontaneous and unscripted DB for authentic emotion recognition. The facial and speech inputs were labeled and marked by many annotators on valence (negative or positive), activation (calm or excited), and dominance (weak or strong). The DB is divided into three modules: VAM-Video, VAM-audio, and VAM-faces.

2.5. Unimodal Databases

2.5.1. AFEW

AFEW is a static and temporal facial expression DB of movie excerpts [169]. This close-to-the-real-world "wild" DB portrays challenging emotional conditions in the clips and is arguably more challenging than most DBs because of the natural head movements, illumination conditions, and the presence of 1+ subject in clips that mimic real-world conditions. A semi-automatic recommender system was utilized to create this DB that

parses the extracted movie subtitles. A human annotator also labeled the clip with the extracted emotional keyword.

The corpus was used to train and test the Emonets, the multimodal DL method for detecting emotions in videos using CNN, DBN, K-means, and relational auto-encoders for classifying the seven emotions and reported an accuracy of 47.67% [5]. AFEW was used together with the static facial expressions in the wild (SFEW) to detect emotions in the wild using a fusion network that combined the audio–visual modalities at the decision level and observed accuracy of 51.02% [2]. A HOG-TOP-based approach for recognizing emotions in the wild integrated audio–visual features using multiple kernel learning (MKL), while an SVM classifier was used for classification and achieved accuracies of 40.21% and 45.21% for the validation set and the test set, respectively [7]. The authors also tested another MER system using a hierarchical classification fusion framework combining voice and facial expressions at the decision level. They used GEMEP-FERA2011 and CK to classify seven emotions with an accuracy of 47.17% [9]. A new geometric feature derived from the wrap transformation of facial landmarks was proposed for capturing facial configuration changes and fused facial features and speech for video-based emotion recognition [16]. AFEW, CK+, and GEMEP-FERA 2011 were used to report the best accuracy of 95.7% (for AFEW), while a combination of HOG-TOP feature and geometric wrap features were used. In another bi-modal emotion recognition system using a sparse kernel reduced rank regression (SKRRR) fusion method with a Gaussian kernel [17], AFEW was used with eNTERFACE'05. The speech features were extracted using the openSMILE feature extractor. The facial features were extracted using a scale-invariant feature transform (SIFT) feature descriptor, and an accuracy of 47% was observed. AFEW was also used to classify seven emotions using an ensemble approach that fused various classification models, SVM and random forest (RF), and reported accuracies of 43.86% and 46.88% for the validation and test dataset, respectively [26].

2.5.2. Berlin

The Berlin speech DB was recorded in German and contained emotional utterances of 10 German (5 female, 5 male) actors, simulated by speaking 10 (5 short, 5 long) sentences [170]. The DB was assessed in an automated listening test and was annotated by 20 listeners to classify the utterances into 7 categories (anger, fear, happy, sad, disgust, boredom, and neutral). The hidden Markov Models (HMM) and ANN approach using the Berlin and CK DB for MER using speech and facial expressions reported an accuracy of 89.6% for decision level fusion [3]. The Berlin speech DB was used with eNTERFACE'05 and CK in an MDR and Ridgelet transform-based approach for audio–visual emotion recognition, where the authors used separate extreme learning machine (ELM) classifiers for each modality and observed an accuracy of 85.13% [14]. The Berlin DB was used with CK to propose an infrastructure for recognizing emotion-aware big data in 5G [15]. Another linguistic and acoustic cues-based bi-modal affective recognition system used ML and valence assessment of the words carrying emotional implications for performing the classification [156]. Berlin speech DB along with SAVEE, polished emotional speech DB [171], electromagnetic articulography (EMA) DB [172], semantic evaluations (SemEVAL) 2007 [173] and international survey on emotion antecedents and reactions (ISEAR) DB [174] was used for testing the accuracy of the proposed approach.

2.5.3. CK/CK+

The Cohn-Kanade (CK) is an AU coded facial DB that was created for emotion recognition research and is available in two versions: CK [175], having data of 97 subjects in 486 excerpts for 7 emotions (anger, contempt, disgust, happy, sad, surprise and neutral) with each peak in the sequence FACS-labeled [176,177], and extended CK (CK+) [146], which adds 593 posed and spontaneous expression excerpts of 123 subjects with validated labels. CK was used for the evaluation of a hierarchical classification framework using a hybrid feature and decision level fusion that used audio–visual features in the wild [9]. This

DB was also used with the acted facial expression in wild (AFEW) DB [169] and GEMEP-FERA2011 to classify seven emotional categories with an accuracy of 47.17%. CK+ was used for a HOG-TOP and geometric wrap-based approach that performed facial expression recognition in the wild using multi-feature fusion and reported an accuracy of 95.7%. In another work, the transfer of emotion features from one DBN to other DBNs was performed with 4 DBs—CK, emoFBVP, Mind reading emotions library and MMI—where CK reported the best accuracy of 89.51% for a training time of 15 h and 49 min [178]. CK+ was also used for a facial expression recognition system incorporating the fusion of base classifiers output with an accuracy of 76.05% [179]. CK along with Berlin DB [170] was used to effectuate a bi-modal emotion recognition using facial expressions and speech [3]. CK was utilized along with Berlin, emoFBVP, DEAP and MAHNOB-HCI databases to perform an unsupervised multimodal emotion classification where accuracy of 97.3% was reported. CK DB was used with eNTERFACE'05 and Berlin to achieve an accuracy of 73.07% for face-only analysis [14]. The authors also used these databases for bi-modal emotion recognition with face and speech using big data and cloud technology and observed an accuracy of 83.10% [15].

2.5.4. FER2013

The FER (facial expression recognition) 2013 DB is a corpus of facial expressions that were produced as a part of a larger project for facial emotion recognition challenge contests [180,181]. A Google search API was used to create the DB using a set of 184 emotion-related keywords. A total of 600 search strings were created for search queries using the emotional keywords and terms related to gender, race, or age for the creation of DB. There are a total of 35,887 images (of size 48 $\times$ 48) for seven emotional categories (anger: 4953 images, disgust: 547 images, fear: 5121 images, happiness: 8989 images, sadness: 6077 images, surprise: 4002 images and neutral: 6198 images) in the DB. There are three variants of FER2013, the FER28, the FER32, and FER32+EmotiW. FER2013 was used in [179] along with JAFFE [182], and CK+ [146] for training and testing an approach based on the ensemble of CNN-based methods, using a probability-based fusion combining all CNN for facial expression recognition. The work classified 7 emotions (neutral, sad, happy, surprise, angry, disgust, and fear) with an accuracy of 69.96%, 69.96% and 66.98% for the FER2013 validation set, FER2013 private set, and FER2013 public set, respectively for ECNN (ensemble convolutional neural network).

2.5.5. JAFFE

The Japanese female facial expressions (JAFFE) DB is a facial corpus of Japanese female models that was created at the Psychology Department, Kyushu University [182–184]. The expressions were posed without any directives and were not tested against any benchmark, while the 240 $\times$ 292 resolution images were labeled by 60 annotators. JAFFE DB was used along with CK+ [146] and FER2013 [180,181] to test a probability-based fusion approach using an ensemble of CNN by combining the outputs of base learning classifiers and reported an accuracy of 50.70% over the CNN (45.07%) [179].

2.5.6. MMI

The MMI (Michel Valstar and Ioannis Patras) DB is one of the most comprehensive and varied facial expression DB available [185,186]. The DB consists of images of 19 subjects of both genders with ages ranging from 19 to 62 years and ethnic backgrounds from Europe, Asia, and South America. The corpus consists of 1500 expression specimens comprising single static images (600 frontal, 140 dual-view, i.e., combining frontal and profile view and recorded using a mirror) and sequences of images (30 profile views and 750 dual-view) in frontal and profiles views. Some part of the corpus (169 sequences) is also labeled with the AU and includes single AU annotation and multiple AU annotation. Despite being encyclopedic, the corpus has some limitations. The DB falls short of spontaneous facial expressions and the annotation of a major part of the corpus with the AU. MMI DB has been used to validate a few systems involving MER. In [178], a layer-by-layer of transfer

of emotion features was proposed in a multimodal setting. The features were transferred between a source (a 6-layer DBN) and a target network (two 6-layer DBN). The authors used three more datasets besides MMI, the emoFBVP [86], Mind Reading Emotions Library, and CK [175] to test the system. An accuracy of 87.39% was reported for the MMI for a training time of 16 h and 38 min.

2.5.7. NTUA

The NTUA (National Taiwan University of Arts) DB is a corpus of emotional utterances that were collected in the Chinese language [167]. The DB was created following the approach of USC CreativeIT [166,167]. A total of 22 exemplars grouped in pairs performed a face-to-face interaction for approximately 3 min in a predetermined emotional scenario. A total of 42 annotators labeled 204 recordings for activation and valence dimensions on a scale of 1–5. NTUA corpus was used for testing and training a cross-lingual emotion recognition system [93] using a multi-task kernel fusion technique. The proposed approach integrated the useful emotional information of one language into another to enhance the affect recognition rate. English language from USC CreativeIT was used with the NTUA to fuse the perspective of the Chinese language for better accuracy. Accuracies of 0.604 and 0.682 was observed in the dimensions of valence and arousal, respectively for the NTUA.

2.5.8. SFEW

The static facial expressions in the wild (SFEW) DB is a facial DB that was created as a subset of AFEW [169,187]. The SFEW holds frontal and non-frontal faces in 700 images with varied head pose age, focus, and illumination for 7 emotions (angry, disgust, surprise, sad, happy, fear, and neutral). SFEW was put to use for testing a fusion network in combining multimodal features for emotion recognition in the wild and reported an accuracy of 51.08% in the classification of 7 emotions [2].

Table 2. Multimodal Databases: A Detailed Comparison.

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
AMIGOS [94]	PS (EEG, GSR and ECG), Video and Depth Information	-	16 emotional movie excerpts (of short duration, <250 s and of long duration, >14 min) in individual and group (4 people) context	40 subjects (13 females and 27 males)	PS recordings in Matlab format: 3.60 GB, Frontal Face Videos: 192 GB, Full body RGB videos from Kinect: 203 GB, Full body depth videos from Kinect: 39.9 GB	Self-assessment by subject of valence, arousal, dominance, liking, familiarity and basic emotions (happy, sad, angry, disgust, surprise, fear and neutral). External annotation of valence and arousal.	[51,188]	Available. The data were recorded in individual settings and group settings to observe and analyze the impact of social context on person's emotional state.	Spontaneous, Induced
ASCERTAIN [97]	PS (EEG, ECG and GSR), Facial activity	-	Emotions induced by watching 36 movie excerpts (ranging from 51–128 s)	58 subjects (21 females and 37 males)	-	Self reports (for 36 videos for 58 subjects each): Arousal, Valence, Engagement, Liking, Familiarity; Personality scores for the Big 5 Personality traits: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness	[189]	Available. First database to connect personality traits and emotional states via physiological responses	Spontaneous, Induced
CALLAS Expressivity Corpus [98,99]	Speech, Facial expressions, Gestures	German Italian Greek	Set of 120 emotion instigating sentences were displayed for the users	21 subjects (10 females and 11 males)	Approximately 5 h of recording	3 emotion categories, positive, neutral and negative	[28]	Unreleased for public domain.DB created to understand nonverbal behaviors across different cultures.	Spontaneous, Induced
CLAS [100]	PS (ECG, EDA, PPG), 3D AD, metadata of the 62 participants	-	Test subjects were involved in some purposely designed interactive or perceptive task	62 subjects (17 females and 45 males)	30-minute recording of physiological signals for each subject. Compressed size: 2.03 GB, Uncompressed size: 13.9 GB	Low/high Arousal, Valence, High-Arousal Negative Valence (HANV) condition, and low/high Concentration.	[190]	ALB. Interactive tasks: solving sequences of math problems, Logic problems, and the Stroop test. Perceptive tasks: use of images and audio–video stimuli for eliciting emotions in the four quadrants of the arousal valence space	Spontaneous, Induced

Table 2. *Cont.*

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
CMU-MOSEI [103]	language (spoken text), visual features (gestures and facial expressions) and audio features (intonations and prosody)	English	-	1000 speakers (43% females and 57% males)	3228 videos acquired from YouTube Total Video hours: 65 h, 53 min and 36 s	Each sentence annotation for a sentiment on a [−3, 3] Likert scale of: [−3: highly negative, −2 negative, −1 weakly negative, 0 neutral, +1 weakly positive, +2 positive, +3 highly positive]. Ekman emotions of happiness, sadness, anger, fear, disgust, and surprise annotated on a [0, 3] Likert scale for the presence of emotion x: [0: no evidence of x, 1: weakly x, 2: x, 3: highly x].	[191]	Available. The corpus consist of 23,453 sentences, 250 distinct topics and 447,143 words.	Spontaneous, Natural
DEAP [104]	EEG, PS, Face	-	Each contributor watched 40 excerpts of one-minute duration music videos and the outcome was recorded in terms of their responses.	32 subjects	Online ratings, video list, participant ratings and participant questionnaire (.ods, .xls and .csv formats). PS: Original BioSemi (.bdf) format (5.8 Gb). Face videos: avi format (15.3 Gb)	Intensity of arousal, valence liking for the video, dominance and familiarity with the videos on a continuous scale of 1–9. a discrete rating of 1–5 was used for appraising the familiarity	[34,36,86]	Available. Out of the total 32 participants, the face video of 22 were recorded.	Spontaneous, Induced
DECAF [105]	BS (204 MEG gradiometers, 102 MEG magnetometers), Face(NIR), PS (3-channels ECG, B-hEOG, B-tEMG)	-	40 music videos as employed in DEAP DB [104], 36 Hollywood movie clips	30 subjects (14 females and 16 males)	14 GB of preprocessed data	Valence–arousal ratings by each participant	[192]	Available. DECAF also contains time continuous emotion annotations for movie clips from seven users	Spontaneous, Induced
emoFBVP [86]	Face, Speech, Body Gesture, Physiological signals	English	Professional actors acted the 23 emotions categories and filled and evaluation form to rate their confidence level with expressing each emotion on a scale of 1 to 5.	10 professional actors	6 recordings (3 in seated and 3 in standing positions) for each actor enacting for 23 different emotions.	23 emotions in 3 intensities (Happy, Sad, Anger, Disgust, Fear, Surprise, Boredom, Interest, Agreement, Disagreement, Neutral, Pride, Shame, Triumphant, Defeat, Sympathy, Antipathy, Admiration, Concentration, Anxiety, Frustration, Content and Contempt).	[86,178]	Not yet released. The actors rated themselves for the acted emotion categories. Paper says database is freely available but website displays database is coming soon!	Posed, Induced

Table 2. *Cont.*

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
EU Emotion Stimulus [106]	Face, Speech, Body gestures, Contextual social scenes	English Swedish Hebrew	Facial expression specific scenes: 249 clips Body gesture specific scenes: 82 clips contextual social science specific scene: 87 clips Speech stimuli: 2364 recordings	19 professional actors	418 visual clips (2–52 s)	21 emotions/mental states: neutral, afraid, angry, ashamed, bored, disappointed, disgusted, excited, frustrated, happy, hurt,interested, jealous, joking, kind, proud, sad, sneaky, surprised,unfriendly, and worried.	[10]	Available. Presents emotional and mental state data for face, body gestures and social sciences contextual scenes not available previously.	Spontaneous, Induced
HEU [107]	HEU-part 1: Face and Body Gesture HEU-part 2: (Face, Body Gesture, and Speech)	Chinese, English, Thai, Korean	-	Total: 9951 subjects (HEU-part1: 8984, HEU-part2: 967)	19,004 video clips	10 emotions: Anger, bored, confused, disappointed, disgust, fear, happy, neutral, sad, surprise	-	Available. Consists of multi-view postures of face and body, multiple local occlusion and illumination and expression intensity.	Posed, Induced
MAHNOB-HCI [85]	32 Channel EEG PS, Face, Audio, EM, Body movements	English	Participants were shown emotional videos and pictures	27 subjects (16 females and 11 males)	Emotion Elicitation exp: 4 readings each for 27 participants, Implicit tagging exp: 3 readings each for 27 participants	Emotional keyworker, Arousal, Valence and Predictability on a Likert scale(1–9)	[86,108]	Available. Subjects from different cultural backgrounds.	Spontaneous, Induced
MELD [109]	Video, Audio, Text	English	-	Characters from TV show "Friends" (84% of time, there were 6 Actors)	13,000 utterances from 1433 dialogues. Each utterance is 3.59 s	7 emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise and Fear; 3 sentiments: positive, negative and neutral	[111]	Available. Each utterance (combination of audio, video and text) is annotated with sentiment and emotion labels.	Posed, Induced
MMSE/BP4D+ [112]	High-resolution 3D model sequences, 2D RGB videos ,Thermal videos, PS(RR, BP, EDA, and HR)	-	Participants conducted 10 activities, under the supervision of a professional actor. The activities included 4 methods: cold pressor, designed physical experiences, interpersonal conversations and watching of a film clip.	140 subjects (82 females and 58 male)	10TB high quality data	10 Emotions: Happiness/amusement, surprise, sadness, startle surprise, skeptical, embarrassment, fear nervous, physical pain, angry, disgust.	[113,114]	Available. Ethnic/Racial Ancestries include Black, White, Asian (including East Asian and Middle East Asian), Hispanic/Latino, and others (e.g., Native American).	Spontaneous, Induced

Table 2. *Cont.*

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
MPED [115]	PS(EEG, GSR, RR, ECG).	-	Each participant watched 28 video clips describing seven different emotions.	23 Chinese volunteers (13 females and 10 males)	5684 s	7 emotions: Joy, funny, anger, disgust, fear, sad and neutrality	[116]	Available. 28 videos for elicitation were manually annotated and were chosen from a group of 1500 videos to reduce the influence of culture	Spontaneous, Induced
MUStARD [117]	Video, Audio, Text	English	-	Characters from 4 TV shows: "Friends", "The Golden Girls", "The Big Bang Theory" and "Sarcasmaholics Anonymous"	690 one-utterance videos	Sarcasm and non-sarcasm	[119,120]	Available. Each audiovisual utterance goes along with the conversational context for additional information.	Posed, Induced
NNIME [121]	Audio, Video, ECG	Chinese	Dyadic spoken interaction between subjects designed by a professional director	44 subjects (22 females, 20 males)	11 h of audio, video, and electrocardiogram data	6 emotions: Angry, happy, sad, neutral, disappointed, surprise	[122]	Available. Consist of both discrete (6 emotion categories) and continuous (activation and valence) emotion annotation	Posed, Natural
RAMAS [123]	Face, Speech, Motion-capture, PS (EDA and PG)	Russian	The subjects actors played out interactive dyadic scenarios involving each of the six emotional categories.	Ten semi-professional actors (5 females and 5 males)	Approximately 7 h of high quality Videos	6 emotions: angry, disgust, happy, sad, scared, surprised; 2 social interaction characteristics: domination and submission	[125]	ALB	Posed, Induced
RECOLA [126]	Speech, Face, ECG, EDA	French	23 teams of two were asked to complete a collaborative task involving social and emotional behaviors.	46 subjects (27 females and 19 males)	9.5 h of audio, visual, and physiological (electrocardiogram, and electrodermal activity) recording.	Continuous dimensions of valence and arousal.	[8,35,127,193]	Available. 34 participants out of 46 gave their consent to share their data.	Spontaneous, Induced
SLADE [128]	PS (EEG, ECG, GSR and ST)	-	Audio-visual stimuli in the form of 40 movie excerpts (duration 1 min)	-	-	Valence-arousal ratings by each participant	-	Not publicly available. This DB was developed to facilitate the detection of stress levels end ER.	Spontaneous/ Induced
AVEC [129,130]	Face, Speech	German	Subjects performed HCI tasks while they were being recorded	292 subjects	340 video clips adding up to 240 h	The clips were annotated in 3 dimensions; arousal, valence and dominance, in continuous time and value	[20,131]	Not available publicly. Developed for emotion and depression recognition challenge.	Spontaneous, Induced

Table 2. *Cont.*

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
BAUM-1 [132]	Face, Speech	Turkish	Subjects watched images and video sequences to evoke target emotions.	31 subjects (17 females and 14 males)	1222 video clips	6 emotions: happiness, anger, sadness, disgust, fear and surprise Intensity on a scale 0–5	[23]	Available. Consists of emotional and mental states. Close to natural conditions such as background.	Spontaneous, Induced
BAUM-2 [194]	Face, Speech	English Turkish	-	286 subjects (118 females and 168 males)	1047 video clips	7 emotions: anger, happiness, sadness, disgust, surprise, neutral and fear. Intensity of emotion in term of score between 1 and 5	-	Available. Contains clips from movies and TV clips are more naturalistic. It is a multilingual database	Posed, Natural
BioVid Emo [133]	PPS (SCL, ECG, tEMG), Video	-	15 standardized film clips of length 32 to 245 s were shown to the participants	94 subjects (50 females and 44 males)	430 .csv sheets; 5 sheets for each subject; each sheet contains PPS data of the participant's chosen film clip 430 .mp4 videos of the .csv sheets	5 emotions: amusement, sadness, anger, disgust and fear.	[134]	ALB. Originally 94 subjects were there but now only 86 remain due to missed or corrupted recordings.	Spontaneous, Induced
CHEAVD [135]	Audio, Video	Chinese	-	238 speakers (47.5% female and 52.5% male)	140 min of excerpts from 34 films, 2 television series, 2 television shows, 1 impromptu speech and 1 talk show	26 non-prototypical emotional states, including the basic six: Angry, happy, sad, neutral, disgust, surprise	[137,138]	Available. Multi-emotion labels and fake/suppressed emotion labels. Highly Skewed	Posed/Induced
CHEAVD 2.0 [136]	Audio, Video	Chinese	-	527 speakers (41.6% female and 58.4% male)	474 min of excerpts from Chinese movies, soap operas and TV shows, 7030 samples	8 emotions: neutral, angry, happy, sad, surprised, disgust, worried, anxious	[137,138]	Available. This DB was developed for the Multimodal Emotion Recognition Challenge (MEC) 2017	Posed/Induced
DREAMER [139]	EEG & ECG	-	18 movie clips (duration: 65–393 s) to elicit emotions	23 subjects (9 females and 14 males)	-	Valence/arousal rated using a discrete scale of integers from 1 to 5	[195,196]	Available. Signals were captured using portable, wearable, wireless, low-cost, and off-the-shelf equipment.	Spontaneous/ Induced
eNTERFACE'05 [143]	Face, Speech	English	Subjects were asked to hear six short stories evoking a particular emotion	42 subjects (8 females and 34 males)	Total: 1166 videos (Anger: 200, Disgust: 189, Fear: 187, Happiness: 205, Sadness: 195, Surprise: 190)	6 emotions: happiness, sadness, surprise, anger, disgust and fear.	[11,14,15,17,21–23,27]	Available. Subjects were from 14 different nationalities (Belgium, Turkey, France, Spain, Greece, Austria, Italy, Cuba, Slovakia, Brazil, USA, Croatia, Canada, and Russia),	Spontaneous, Induced

Table 2. *Cont.*

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
FABO [144]	Face Expressions Body Gestures	-	Subjects provided with situation or short scenarios for narrating an emotion eliciting situation.	23 subjects (12 females and 11 males)	Video size of 9 Gb	10 emotions: neutral, uncertainty, anger, surprise, fear, anxiety, happiness, disgust, boredom, sadness	[145]	Available. DB consists of subjects from different ethnicities, such as European, Middle Eastern, Latin American, Asian, and Australian	Posed, Induced
GEMEP-FERA 2011 [147,148]	Face, Speech	2 pseudo-linguistic phoneme sequences.	Actors were guided by a professional director and uttered 2 pseudo-linguistic phoneme sequences or a sustained vowel 'aaa'	10 professional actors (5 female and 5 male)	7000 audio-video illustrations	18 emotions (admiration, amusement, anger, tenderness, disgust, despair, pride, shame, anxiety, interest, irritation, joy, contempt, fear, pleasure–sensual, relief, surprise and sadness)	[16]	Not available publicly. Very subtle and not very common 18 emotional categories.	Posed, Induced
HUMAINE [149]	Face, Speech, Gestures	English, French, Hebrew	-	-	50 video clips ranging from 5 s to 3 min	Emotional content at 2 levels: Global label, applied to whole clip, continuous annotation on one dimensional axis such as valence, activation, arousal, intensity or power.	[25]	Authors mentioned available, but the link is broken. Contains clips from TV chat shows and religious shows data annotated for audio visual only.	Posed, Natural
IEMOCAP [93]	Face, Speech, Head and Hand movement, DT, Word, Syllable and Phoneme level alignment	English	Subjects enacted in pairs on 3 selected emotional scripts in hypothetical scenarios with markers on their face, head and hands.	10 professional actors (5 female and 5 male)	Approximately 12 h of data	5 emotions in beginning (happiness, anger, sadness, frustration, neutral). Later (surprise, fear, disgust, excited) were added.	[13,19,33,131]	Available. Annotation done for utterances by at least 3 annotators annotated on valence, activation and dominance.	Spontaneous, Induced
LIRIS-ACCEDE [150,151]	Audio, Video	English, Italian, Spanish, French	-	-	9800 high quality movie excerpts (duration 8–12 s) taken from 160 movies	Rankings for arousal and valence dimensions	[197]	Available. The DB was annotated by 1517 trusted annotators from 89 countries.	Spontaneous, Natural
RAVDESS [152]	Audio, Video	English	Two neutral statements were used ("Kids are talking by the door", "Dogs are sitting by the door")	24 professional actors (12 females and 12 males).	7356 recordings (total size: 24.8 GB)	8 emotional categories: neutral, calm, angry, sad, fearful, surprise, disgust and happy; songs: neutral, sad, angry, calm, fearful and happy	[18,114,153,154]	Available. Each expression is produced at two levels of emotional intensity (normal, strong)	Posed, Induced

Table 2. *Cont.*

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
RML [155]	Face, Speech	English Mandarin Urdu Punjabi Persian Italian	10 different sentences for each emotional class	8 subjects	720 visual clips of 5 s each. Total Size: 4.2 Gb	6 emotions: (anger, disgust, fear, happiness, sadness and surprise)	[18,23,156]	Available. Participants were asked to recall any emotion related event in their lives to exhibit natural emotion	Spontaneous, Natural
SAL [149,157]	Face, Speech	English Hebrew Greek	Conversation between a machine and human. Machine had emotional traits	20 subjects	491 recordings total of approx. 10 h	Four emotions: happy, gloomy, angry and pragmatic	-	Available. Low intensity of recorded emotions.	Spontaneous, Induced
SAVEE [38,158]	Face, Speech	British English	Emotional video clips, texts and pictures were displayed for the participants (3 for each emotion)	4 native English males	480 recordings (120 for each subject)	Seven emotions: (anger, disgust, fear, happiness, sadness, surprise and neutral)	[19,24,156]	Available.Evaluated by 10 evaluators for accuracy	Spontaneous, Induced
SEED [159–161]	EEG	-	15 Chinese films clips of 4 min	15 subjects (8 females and 7 males)	43.88 GB	3 emotional states: neutral, happy and sad)	[160]	Available.Consists of EEG and Eye movement for 12 subjects and only EEG Data for 3 subjects.	Spontaneous, Induced
SEED-IV [162]	EEG and EM	-	72 Video clips	15 subjects (8 females and 7 males)	6.88 GB	4 emotional states: Happy, sad, fear, and neutral	[160]	Available.	Spontaneous, Induced
SEED-V [163]	EEG and EM	-	15 movie clips	16 subjects (10 females and 6 males)	37.77 GB	5 emotional states: happy, neutral, disgust, sad and fear	[160]	Available.	Spontaneous, Induced
SEED-VIG [164]	EEG and EOG	-	A 4-lane highway scene, a VR-based driving system	23 subjects (12 females and 11 males)	2.94 GB	Vigilance labels (ranging from 0 to 1)	[160]	Available.	Spontaneous, Induced
SEMAINE [165,198]	Face, Speech	English	Subjects talked to a SAL (sensitive artificial listener)	150 subjects	959 conversations of around 5 min duration of each clip.	5 affective dimensions (arousal, expectation, intensity, power and valence) for 27 categories	-	Available. Participants from 8 different countries.	Spontaneous, Induced
The USC CreativeIT [166,167]	Speech, Body movements	English	Subjects in pairs did a conversation producing expressive affective behaviors.	16 actors (8 females and 8 males)	90 recordings of 3 min each	Annotated on Valence, activation and arousal on a scale 1–5	[167]	Available.	Spontaneous, Induced

Table 2. *Cont.*

Database	Modalities	Language	Elicitation Method	Subjects	Size of Databases/No. of Samples	Emotion Description	Representation Work	Remarks	Posed/Spontaneous Induced/Natural
VAM [168]	Face, Speech	German	Extracted from German talk show "Vera am Mittag"	VAM-video 104, VAM-audio 47, VAM-faces 20 subjects	VAM-video 1421 utterances VAM-audio 1018 utterances VAM-faces 1872 images	Emotion marked on 3 continuous emotion parameters, i.e., valence (negative or positive), activation (calm or excited) and dominance (weak or strong).	-	Available.	Spontaneous, Natural
AFEW [169]	Face	-	-	330 subjects	957 video clips ranging from 300 ms to 5400 ms	Six emotions:angry, disgust, fear, happy, sad and surprise	[2,5,7,9,16,17,26]	Available. Referred to as wild because close to the real world.	Spontaneous, Natural
Berlin [170]	Speech	German	Simulated the emotion by speaking 10 sentences	10 actors (5 females and 5 males)	800 sentences	6 emotions: anger, fear, happy, sad, disgust and boredom	[3,14,15,156]	Available. Each utterance was judged by 20 listeners.	Posed, Natural
CK/CK+ [146,175]	Face	-	Subjects were told to perform facial displays by the instructor	CK: 97 subjects, CK+: 123 subjects	CK: 486 excerpts CK+: 593 excerpts	7 emotions: anger, contempt, disgust, happy, sad, surprise and neutral	[3,9,14–16,86,178,179]	Available.Another version of the Database is being planned for future release.	CK: Posed CK+: Posed and spontaneous, Natural
FER2013 [180,181]	Face	-	-	-	35,887 images of size 48 × 48	Seven emotions: anger, disgust, fear, happy, sad, surprise and neutral	[179]	Available. 3 variants of database: FER28, FER32 and FER32+EmotiW.	Spontaneous, Natural
JAFFE [182–184]	Face	-	Expressions were posed without any directives	10 subjects	217 images	7 emotions: happy, sad, fear, anger, surprise, disgust and neutral	[179]	Available. 60 Japanese annotators labeled images. Size of image 240 × 292	Posed, Natural
MMI [86]	Face	-	Each participant was asked to display all 31 AUs and a number of extra Action Descriptors.	Phase I: 19 (8 females and 11 males) PhaseII: 75	Phase I:740 images 848 videos PhaseII: 2900 videos and high resolution images	169 sequences are FACS coded. The DB comprise temporal pattern of expression: neutral-onset/apex/offset-neutral	[178]	Available. Frontal and dual view (frontal with profile view)	Posed, Natural
NTUA [167]	Speech	Chinese	Dyadic interaction between subjects designed by a professional director	22 subjects	204 recordings (approx. 3 min each)	Emotion labeled in arousal and valence dimension on a scale between (1,5)	[167]	Availability: unknown. Annotated by 42 annotators	Posed, Natural
SFEW [187]	Face	-	-	95 subjects	700 images	Seven emotions: angry, disgust, surprise, sad, happy, fear and neutral	[9]	Available. Subset of AFEW [169]	Spontaneous, Natural

3. A Taxonomy for the Recognized Features in Modalities

Section 2 presents the description of the DBs involved in MER. The modalities presented in the DBs for MER provide multiple ways to identify the emotional state of a person. The section also presents different ways of utilizing these DBs through various ML and DL methods. These methods rely on robust and high-quality features for performing MER. This section plunges into more granularity by presenting the features and the taxonomy of the features identified and used by the methods mentioned in Section 2.

Figure 2 presents the proposed taxonomy of the recognized features in the observed modalities of emotion recognition. Tables 3–5 define the features identified in physiological signals, speech, and facial expressions, respectively. These features could not be accommodated in Figure 2 due to lack of space and therefore forms a reference for the figure. The recognized features form an important component of emotion recognition regardless of the modalities or the combination of modalities. The modalities that were identified are as follows:

- Facial Expressions
- Physiological Signals
- Speech
- IR/ Thermal Images
- Contextual Social Science Scenes
- Eye Gaze
- Body movements/ Gestures
- Dialogue Transcriptions
- Head movements and Head angle information
- Word-level, syllable level and phoneme level alignment

Speech, physiological signals, and facial expressions were the top modalities with the highest number of features in the same order. Other modalities were not so varied in handcrafted features, while some did not mention the features utilized for affect recognition.

Table 3. Speech features.

Pitch	S01	Entropy	S36	RASTA based on PLP	S71
Signal	S02	Slope	S37	f0 mean value	S72
Energy	S03	Psychoacoustic Sharpness	S38	25% and 75% quantiles	S73
MFCC	S04	Harmonicity	S39	difference between f0 max and min	S74
Spectral	S05	Spectral Variance	S40	difference of quartiles	S75
Voice quality	S06	Skewness	S41	unvoiced and voiced segments duration ratio	S76
Mean	S07	Kurtosis	S42	avg duration of voiced segments	S77
Median	S08	F0 (SHS & Viterbi smoothing)	S43	intensity mean	S78
Max	S09	Prob. of voicing	S44	energy range	S79
Min	S10	log. HNR	S45	avg pause length	S80
Variance	S11	Jitter (local &)	S46	speaking rate	S81
Lower/upper quartile	S12	Shimmer (local)	S47	min and max of Mel freq	S82
Absolute/quartile	S13	27 Mel Freq Band coef (MFB)	S48	mean and std dev of I and II Gaussian of MFCC	S83
Harmonics-to-noise ratio	S14	Pitch mean	S49	min and max of Mel freq first order difference	S84
Mel-freq filter banks	S15	Pitch median	S50	mean and std dev of Mel freq I order difference	S85
MFCC Derivative	S16	pitch std dev	S51	mean and std dev of total log energy; mean	S86
MFCC Autocorrelation	S17	std dev of duration of voiced segments	S52	std dev	S87
Formants (Formants up to 5500 Hz)	S18	pitch range	S53	voiced speech duration	S88
Time freq	S19	pitch variation rate	S54	unvoiced speech duration	S89
MSpectrum flux	S20	rising/falling ratio	S55	sentence duration	S90
Spectral centroid	S21	rising pitch slope max	S56	average voiced phone duration	S91
Delta spectrum magnitude	S22	falling pitch slope map	S57	average unvoiced phone duration	S92
Band energy ratio	S23	rising pitch slope mean	S58	voiced-to-unvoiced speech duration ratio	S93
Zero crossing rate (Avg, Std dev)	S24	falling pitch slope mean	S59	avg voiced-to-unvoiced speech duration ratio	S94
Silence ratio (Prop of silence in a time window)	S25	pitch rising range max	S60	speech rate (phone/s)	S95
pitch max	S26	pitch falling range max	S61	voiced-speech-to-sentence duration ratio	S96
Sum of auditory spectrum (loudness)	S27	pitch rising range mean	S62	unvoiced-speech-to-sentence duration ratio	S97

Table 3. *Cont.*

Sum of RASTA-filtered auditory spectrum	S28	pitch falling range mean	S63	32 energy and spectral related LLD × 42 func	S98
RMS Energy	S29	overall pitch slope mean	S64	6 voicing related LLD × 32 func	S99
Zero-Crossing Rate	S30	overall pitch slope std dev	S65	32 delta coef of energy/spectral LLD × 19 func	S100
RASTA-filt. aud. spect. bds. 1–26 (0–8 kHz)	S31	overall pitch slope median	S66	6 delta coef of voicing related LLD × 19 func	S101
Spectral energy 250–650 Hz, 1 k–4 kHz,	S32	energy mean	S67	avg duration of unvoiced segments	S102
Spectral Roll-O Pt. 0.25, 0.5, 0.75, 0.9	S33	energy std dev	S68	std dev of duration of voiced segments	S103
Spectral Flux	S34	energy max	S69		
Centroid	S35	10 voiced/unvoiced durational features	S70		

Legend: RASTA: relative spectral features, PLP: perceptual linear prediction, avg: average, freq: frequency, -ve: negative, LLD: low level descriptors, coef: coefficients, prop: proportion, std: standard, dev: deviation, func: functionals, MFCC: Mel Frequency Cepstral Coefficients.

Table 4. Facial expression features.

A (HL-inner eye corners, L-inner and outer eyebrow)	F01	Lip pucker	F22	SPTS	F43
VD (outer eyebrows, L bw inner corners of eyes)	F02	Lips part	F23	CAPP	F44
D (outer eyes' corner, their upper eyelids)	F03	D(lower lip and mouth corners)	F24	PHOG	F45
D (inner eyes' corner, their upper eyelid)	F04	D (between mouth corners)	F25	LBP-TOP	F46
D (outer eyes' corner, their lower eyelids)	F05	vertical D (upper and the lower lip)	F26	opening of mouth	F47
D (inner eyes' corner, their lower eyelids)	F06	Bounding box of face	F27	facial expression	F48
V D (upper eyelids, the lower eyelids)	F07	position of eyes	F28	Mean	F49
D (upper lip, mouth corners)	F08	position of mouth	F29	Energy	F50
15 facial AU	F09	position of nose	F30	Std dev	F51
head-pose in 3D	F10	Jaw drop	F31	Min	F52
mean and std dev (optical flow in head region)	F11	Local Binary Patterns (LBP)	F32	Max	F53
Gabor wavelet features	F12	240 VF (2D MC as mean and std dev of adjusted MC)	F33	Range	F54
Inner brow raiser	F13	Local Phase Quantization (LPQ)	F34	position min/max	F55
Outer brow raiser	F14	Patterns of oriented edge magnitudes (POEM)	F35	number crossings/peaks	F56
Brow lowerer	F15	LPQ-45	F36	length	F57
Cheek raiser	F16	PHOG-46	F37	Geometric	F58
Lid tightener	F17	PHOG-45-PCA	F38	2D global head motion estimation	F59
Upper lip raiser	F18	LPQ-all	F39	A measure	F60
Lip corner puller	F19	LPQ-all-PCA	F40	two-rectangle features	F61
Lip corner depressor	F20	PHOG-all	F41	three-rectangle features	F62
Chin raiser	F21	PHOG-all-PCA	F42	four-rectangle features	F63
		AU 1–2, 4–12, 14–18, 20, 22–28, 38–41, 42–46	F64		

Legend: AU: Action Units, avg: average, A: angles, D: distance, VF: visual features, coef: coefficients, MC: marker coordinates, stddev: standard deviation, H: horizontal, L: line, V: vertical, bw: between, PCA: principal component analysis, PHOG: pyramid histograms of oriented gradients, LBP-TOP: local binary pattern from three orthogonal planes.

Table 5. Physiological signal features.

avg skin resistance	P01	E ratio bw the freq bands [0.04–0.15] Hz, [0.15–0.5] Hz	P14	median peak to peak time	P27
avg of D	P02	SP in bands ([0.1–0.2] Hz, [0.2–0.3] Hz, [0.3–0.4] Hz)	P15	avg	P28
avg decrease rate in decay time	P03	low freq [0.01–0.08] Hz	P16	avg of D	P29
prop of -ve samples in the D vs. all samples	P04	medium freq [0.08–0.15] Hz	P17	Band SP in ([0–0.1] Hz, [0.1–0.2] Hz)	P30
no. of local minima in the GSR signal	P05	high freq [0.15–0.5] Hz	P18	EMG	P31
avg rising time of the GSR signal	P06	ECG (64)	P19	eye blinking rate	P32
10 SP in the [0–2.4] Hz bands	P07	band E ratio log(E(0.05–0.25 Hz))-log(E(0.25–5 Hz))	P20	E of the signal	P33
ZCR of SCSR [0–0.2] Hz	P08	avg respiration signal	P21	mean and variance of the signal	P34
ZCR of SCVSR [0–0.08] Hz	P09	mean of the respiration signal	P22	10 SP in the bands from 0 to 2.4 Hz	P35
SCSR	P10	std deviation	P23	avg peak to peak time	P36
SCVSR mean of peaks magnitude	P11	range or greatest breath	P24	Heart Rate Variability (HRV)	P37
avg and std deviation of HR	P12	breathing rhythm (spectral centroid)	P25	inter beat intervals	P38
breathing rate	P13	inter beat intervals	P26		

Legend: avg: average, D: derivative, E: energy, freq: frequency, -ve: negative, SP: spectral power, cond: conductance, prop: proportion, std: standard, SCSR: skin conductance slow response, SCVSR: skin conductance very slow response, ZCR: zero crossing rate, bw: between.

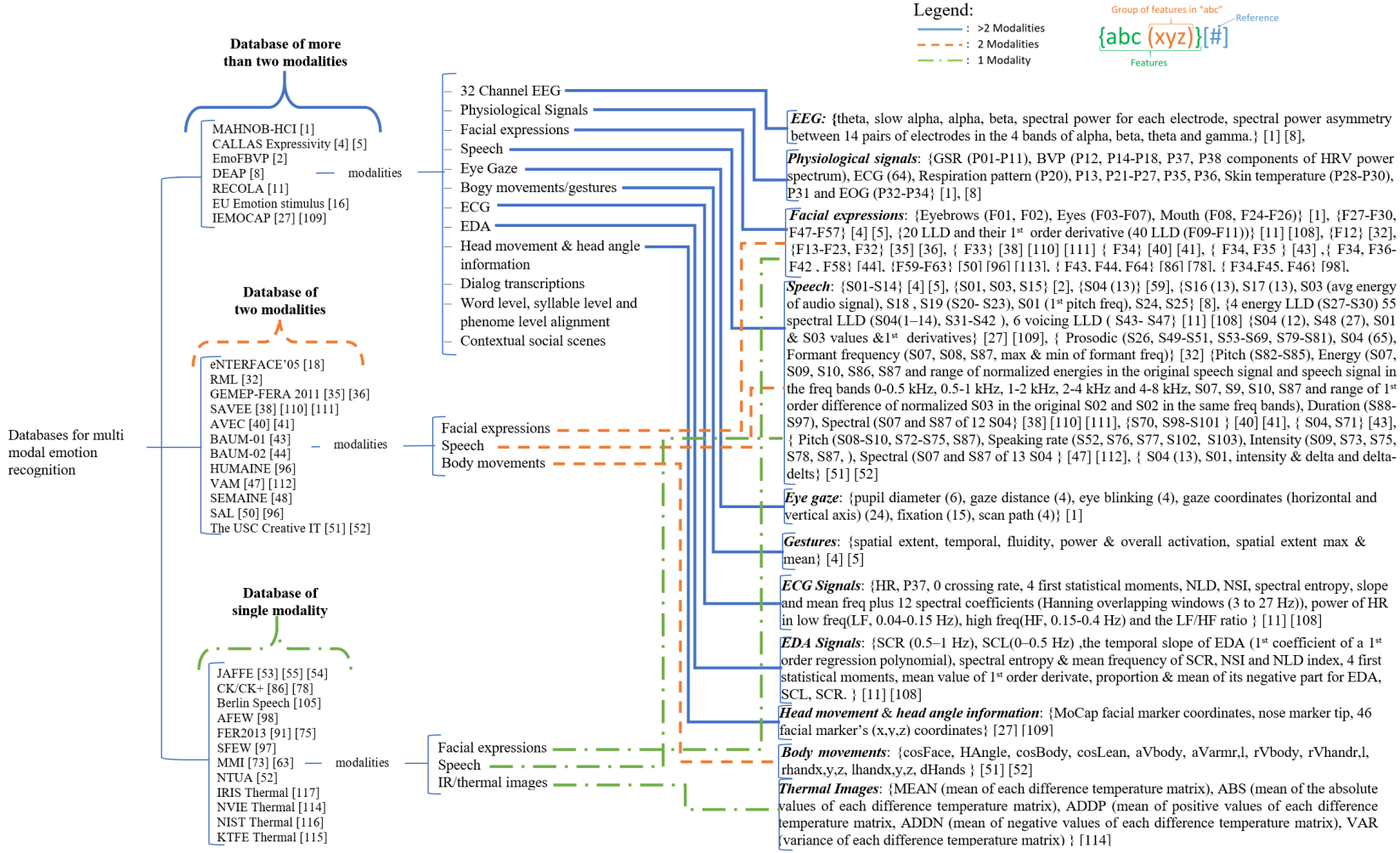


Figure 2. Taxonomy of the features applied in MER. Acronyms: BVP, blood volume pressure; NLD, normalized length density; NSI, non-stationary index.

4. The VIRI DB

4.1. Motivation

Another type of input forms a relatively untapped member of the kins for emotional recognition: the IR images. Not used much in recent times, the IR images seem a propitious route for experimentation in affect recognition. Their ability to retain ample information even in low lighting conditions supports their conspicuous nature. Scant DBs of the IR images do exist in the literature and are available to be used for research. One IR facial DB, IRIS [89], is a corpus containing thermal and visible face images under variable illumination. Recorded with a resolution of 320 × 240, the IRIS dataset consists of 30 probands (28 males, 2 females) expressing three different emotions (surprise, laugh, anger). There are 176–250 images for every person, and 11 images/rotations were recorded for each subject. Providing 4228 pairs of thermal and visible images, the DB provides 5 illumination conditions while utilizing the rotary system. A defunct DB available earlier, the NIST [90], was also used in a few pieces of research involving thermal facial images. The DB was for three posed expressions (smile, frowning, and surprise) and contained 1919 IR images for more than 600 subjects. A third DB in this category, the NVIE [91], provided images in both posed and spontaneous categories in IR and visible format. There were three illumination conditions (right, left, and front) in which the images were shot for 215 individuals (157 males and 58 females), ages ranging from 17 to 31 years. Videos clips downloaded from the internet were used to invoke desired emotions into the subject. The expressions were recorded for both induced emotions (by the videos) and the six acted emotions (happiness, sadness, surprise, fear, anger, and disgust). Another thermal emotion DB that contains spontaneous expressions for the six emotional categories (happiness, sadness, surprise, fear, anger, and disgust), along with the neutral, is the KTFE [92]. The corpus is ethnically rich, containing visual and thermal videos for 26 subjects from Vietnam, Japan, and Thailand in the age groups of 11 to 31 years. Stimuli in the form of emotional excerpts were used to invoke emotions in the probands, similar to the procedure carried out to create the NVIE dataset. However, the KTFE dataset is further classified into four categories for the intensity levels of the induced emotions.

The thermal emotion DBs are rare, and the existing DBs have some advantages and disadvantages over each other. Table 6 compares the existing DB with the proposed VIRI DB. The creation of a DB, despite the availability of an existing one, is to address the limitation posed by the latter. If the IRIS and NIST DB had fewer emotional categories (just 3 in each case), the NVIE DB lacked in all 6 emotional expressions triggered spontaneously. The KTFE DB, how quintessential it is, has all subjects recorded in controlled laboratory conditions. The emotion recognition in uncontrolled environments requires a DB that contains training data captured in the wild with real-life-like conditions. To address the limitations in the existing thermal DBs for emotional recognition, an up-to-the-minute DB, the VIRI (visible images and IR images) DB is proposed to facilitate pragmatic research in emotion recognition. The following few sections give an overview of the material and methods hired to create the corpus.

Table 6. Comparison of the available IR DBs and VIRI.

IR DB	# Participants	Emotions	WB	S/P	Available
IRIS	30 (28 M, 2 F)	S, L, A	No	P	Yes
NIST	600	SM, F, S	No	P	No
NVIE	215 (157 M, 58 F)	H, SA, S, FE, A, D, N	No	S, P	Yes
KTFE	26 (16 M, 10 F)	H, SA, S, FE, A, D, N	No	S	Yes
VIRI	110 (70 M, 40 F)	H, SA, S, A, N	Yes	S	Yes

Legend: S: Surprise, L: Laugh, A: Anger, SM: Smile, F: Frown, H: Happy, SA: Sad, FE: Fear, D: Disgust, N: Neutral, WB: Wild Background, S/P: Spontaneous/Posed.

4.2. Process of DB Creation

The VIRI DB was created to provide a convenient corpus with spontaneous facial expressions in both visible and IR format in uncontrolled environments. The research in facial expressions needs to be superintended to real-world-like situations, and this DB will augment the motion. The presented DB was created at The University of Toledo. Five expressions were captured in this DB (happy, sad, angry, surprised, and neutral). This section presents the description of the participants, the devices used, the environments in which the corpus was created, and the procedure that was followed for building the DB.

4.2.1. Procedures and Participants

The corpus was created at the University of Toledo and comprised the campus students. The process has been approved by The Human Research Protection Program & Institutional Review Boards (IRB) of the University of Toledo (# 202741). The students were told the motive of the DB creation and were solicited to obtain their images captured with the expressions. They were asked to sign a consent form before their images were taken. Once they signed the form, they were told or asked particular emotion-eliciting questions that would enable them to present spontaneous emotions based on the content of the question. There was not any specific set of questions. The question, for instance, asked what they would do when "it was announced that Spring break has been canceled", alternatively, what they would do when "they will come to know that they have been selected for a huge lottery". A total of five images with the required five expressions were captured at an interval of 10–15 s between the two consecutive captures. The subjects were asked to remove glasses (as glass absorbs IR radiations) that could prune the information meant to be captured.

A total of 110 subjects participated in the creation of VIRI DB. A majority of the participants were males, with 70 males compared to 40 females. The DB is diverse ethnically, with subjects evenly distributed among Americans, Afro-Americans, and Asians. Most of the participants were students on the campus, and their ages ranged from 17 to 35 years. The task of DB creation took nearly three months. The aim was to create a DB where the subject is in the wild. This was further governed by the fact that consent of the subject was required before the capture. Authors used to ask, and many times, people would say no to the experiment. Also, not all the subjects captured were included in the DB. There were certain cases where the emotions represented were not correct and could act as an outlier that would affect the ML or DL algorithm's accuracy. This is why the DB comprises a small set of 110 participants. However, this is a good number compared to other DBs, where the actual subject was there instead of the DB created from movie or TV show recordings (e.g., AMIGOS: 40 subjects; ASCERTAIN: 50 subjects; CLAS: 62 subjects). Additionally, there was no emphasis on the gender balance of the subjects, and they were selected based on the correctness of the emotion portrayed. Each participant posed for five emotions, namely happy, sad, angry, surprise, and neutral. Some of the subjects could express emotions more distinctly, while some could not.

4.2.2. Devices

The images were captured using a FLIRONE Pro thermal camera (Figure 3) for an android device. The device is manufactured by FLIR Systems, Inc. and can capture visible and IR images simultaneously. The device is highly portable, which aided in taking the images of the subjects on the go. With a thermal resolution of 160 × 120 and a visual resolution of 1440 × 1080, the thermal imaging device is capable of measuring temperature ranges between −4 °F and 752 °F (−20 °C to 400 °C). The spectral range of the device's thermal sensor is 8–14 µm with a pixel size of 12 µm. The device can exhibit up to three spot temperature meters and six temperature regions of interest. Apart from pure thermal and pure visible images, the device is adept at taking a third type of image format, the MSX (multi-spectral dynamic imaging) format. An MSX format is a FLIRONE Pro's feature where the edge detail of the visible image is etched on the thermal image for more detail.

The VIRI DB will comprise these MSX format images for the IR images to incur more detail and information for training the systems of affect recognition. The image format captured through the camera is a radiometric jpeg that contains both a spatial representation of the scene and thermal data that exhibit the radiant energy seen by the camera.

Figure 3. FLIRONE pro thermal camera.

An android device (Samsung Galaxy S8+) was used to bestride the FILRONE Pro thermal camera through a USB C-type port. The device was running on Android version 7.0 (Nougat) with a 4.00 GB RAM. An app from FLIR, named the FLIR ONE version 2.1.25 was used to capture images through a dedicated interface.

4.2.3. Viri DB Design

One of the main reasons behind the creation of VIRI DB creation was the need for spontaneous facial expressions DB in a visible and thermal format that is taken in the wild with no emphasis laid on the clean backgrounds. The presented corpus considers the above conditions and provides a DB of simultaneous visible and thermal images in the wild. The images were not taken in the laboratory conditions with a clean background; instead, care was taken to capture them in uncontrolled environments such as cafeterias, auditoriums, hallways, classes, etc.

A total of 550 images in radiometric jpeg were captured for 110 subjects expressing 5 different emotions. The software, FLIR Tools, provided by FLIR Systems, Inc., was used to extract the visible and MSX Thermal images from the captured radiometric format. This resulted in a total of 1100 images (550 visible and 550 IR images). Figures 4 and 5 show the sample images of the two subjects in the VIRI DB. The resolution of each image was 1440 $\times$ 1080 after extraction due to the interpolation done by the FLIR Tool software. Since the images were captured in the wild, some of the subjects in the images were not centered, and hence a batch process of the crop was performed on the images to bring the subjects to the center. Later, the size of each image was reduced to 500 $\times$ 500 pixels to bring it to a realistic size, suitable for use in training. Since the visible images were captured using the visible image camera of the FLIR ONE Pro without a flash, this resulted in the capture of a few underexposed images. To address this, the images were edited to increase exposure. The DB is available for use by contacting the authors. A more in-depth analysis of the use of this database for emotion detection was presented in our previous work [199] and readers may refer to this work for more details on the achieved results. In summary, the results proved that this newly proposed database provides better results for emotion detection. In addition, we also added a third modality, speech, to further improve the results [199].

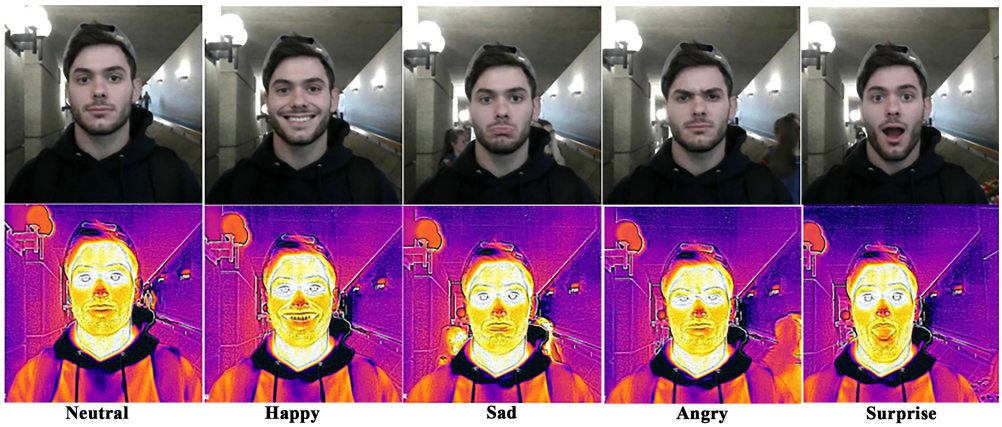


Figure 4. Sample image of a subject showing all five emotions in visible and MSX format in the proposed VIRI DB.

Figure 5. Sample image of a subject for happy and surprise emotions in the proposed VIRI DB.

5. Future Research Directions

Several current and growing areas have applications of multimodal emotion recognition. Some of these are listed as follows:

5.1. Artificial Emotional Intelligence (AEI)

Artificial intelligence (AI) is being denominated as the new *electricity* [200], as it is causing the same change as the introduction of electricity which changed the way the world operated around a century ago. As an intrinsic constituent of AI, AEI is also being tagged as the future of AI [201,202]. Research in this aspect of AI is buoyant and continuously evolving to penetrate the untapped realms where this field could be beneficial.

5.2. Automatic Facial Expression Recognition (AFER) Refinement

Facial expressions boast being the prime source of affect recognition [203] and still form the most indicative, intuitive, and robust form of expressing the psychological state of a person. There are a few limitations that need to be addressed in this aspect of affect recognition.

- Unavailability of comprehensive DBs: The facial expression DBs are limited and are not all inclusive. There is a need to create such DBs containing natural expressions. Another aspect of an ideal DB missing in most of the current DB is the varied lighting conditions and the subject captured in an uncontrolled wild environment.

- Rotation invariance: One of the major drawbacks of the AFER is its inability to cope with head pose variations. This aspect of AFER needs attention to carry out impeccable emotion recognition. A way to resolve this could be using 3D facial data that capture depth information.
- Micro expressions: These are brief expressions that a person exhibits in high-stakes situations while expressing their feelings. Difficult to perceive by a novice person, an automatic recognition of these expressions might reveal the feelings that a person is trying to hide. An apposite demonstration of such an application of emotion recognition is shown in the award-winning television series "Lie to Me" [204] inspired by Dr. Ekman's research.
- Intensity: The intensity of a facial expression may serve more value in analyzing behavior and may serve as a tool for differentiating a posed and spontaneous emotion. This arduous task of identifying the intensities of the observed emotion is still in its nascent stage and requires attention to perceive subtle emotional categories.
- Secondary emotional states: Most AFER are restrained to detecting only primary emotional states, such as happy and sad. Many secondary emotional states, such as frustration, pain, depression, etc., necessitate consideration.

5.3. Contextual Emotion Recognition

Context is the environment in which a person is exhibiting affective behavior. If the system is aware of the context, it becomes easy to perceive an emotional state [205]. The contextual emotional recognition may help discern between what is an expression on the face and the actual emotional state.

5.4. Wireless Emotion Recognition

If a person knows that they are being captured or there are sensors on their body, the naturalness of the emotional behavior would be a lost thing. Envisaging the ways to detect the emotional behavior of a person without their knowledge would furnish a pristine emotional state.

5.5. Biometric Surveillance and Monitoring

Security is crucial in this erratic world, and an emotional state of a person might be employed to strengthen it. Their emotional state evinces a person's intention, and more than often, feelings can be a very effective way to predict actions. If a scared person performs an ATM transaction, the machine might deny the transaction detecting suspicious behavior. Similarly, a biometric surveillance system [206] might be able to capture subtle emotions and could prevent a mishap compared to their human counterparts by sensing the intentions of a person automatically.

5.6. Automated Feedback

Whether to take honest feedback from a class or to get a review of the movies, automated emotion recognition can prove to be more than helpful. Disney is already employing such practices, where facial expression recognition is being employed to predict how a person will react to movies [207].

6. Conclusions and Discussion

Affect recognition is finding gravity in a multitude of disciplines, and the automatic discernment of emotion is a dominant area of research in HCI. Various affective recognition systems based on facial expressions, speech, physiological signals, or body movements are being researched to increase the exactness and reliability of emotion identification.

The paper presents an extensive survey of the multimodal DBs used to validate and test the state-of-the-art methods for affect identification. The study describes DBs in the form of the number and type of modalities included, the description of the subjects, the procedure carried out to create the DBs, and the basis of whether the emotions are acted or

spontaneous. The survey also mentioned the research works and reported the accuracy of the emotion recognition systems where these DBs have been used. The survey was carried out in three sections based on the number of modalities involved, i.e., more than two modalities, two modalities, and one modality.

Several types of modalities were observed in the survey, with some of them being used in several combinations with other input types [50,66,80,83,208–224]. Facial expression images, being the prime modality in affect identification, were found in most DBs. The speech was the second most widely used modality, followed by PS. Several other modalities that were noted in the survey included various types of PS (EEG, ECG, GSR, BVP, respiration pattern, ST, EMG, EOG and EDA), hand gestures, body movements, thermal facial images, haptic physical interaction data, language, and acoustic and lexical features of spoken dialogues. The combination of facial expressions and utterances forms one of the most widely used combinations of modalities. Several works also used body movements or physiological signals and the face and speech.

A primary conclusion that can be derived from the study is that the system's accuracy for affect recognition increased proportionally with the increase in the number of modalities being used in the system, i.e., MER has proven to be more accurate than its unimodal compeer. A wide variety of DBs is being employed to train the models of ML approaches to achieve the desired results for MER systems. DB, such as MAHNOB-HCI [85], CALLAS Expressivity Corpus [98,99], emoFBVP [86] or DEAP [104], are multimodal and autarkic in training multimodal systems. A myriad of bi-modal DB were also created to assist the bi-modal emotion recognition through dyads of modalities such as face and speech, face and body movements, etc. eNTERFACE'05 [143], IEMOCAP [93], RML [155] and GEMEP-FERA 2011 [147,148] are the examples of such DB. Another variety of DB incurred in the development of multimodal affective systems is the combination of many unimodal DB such as CK+ [146] and Berlin DB [170]. These DBs were used in [15] to detect emotions working in conjunction. A few of the DBs consisted of the acted emotion dataset, while few were spontaneous. The majority of the DBs were created in controlled laboratory conditions. However, some DBs stuck to the real-world-like environment and contained images captured in the wild, i.e., in uncontrolled environments. It was observed that most of the DBs in the wild were excerpts from video clips of movies and were more challenging than their laboratory counterparts. A detailed comparison of these databases is provided in Table 2. In addition, a brief classification of databases was also performed based on their application to specific domains in Table 7. This table may further help readers in choosing a database based on their needs.

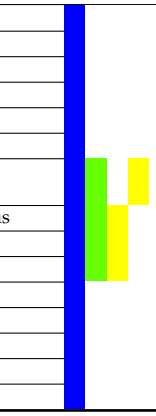
The last section of the paper introduces a new visible image and the IR image DB, the VIRI DB. The DB presents an infrequently used modality: thermal images. A novel thermal image called an MSX format image is used, where the edges of a visible image are embossed over the IR image to give more detail. Very few DBs exist for IR facial expression images. It was noticed that there were certain limitations with the existing corpora. Some of them lacked spontaneity in the captured emotions, while some DBs contained only a few emotion categories. One of the available DBs is already defunct, and all of the DBs are captured in controlled laboratory conditions. A new DB was proposed to address these limitations that contain simultaneous visible and IR images for five emotional categories. The DB is varied in terms of age and ethnicity and contains images taken in the wild.

Table 7. Suitable databases for various application Domains.

Application Domain	Databases and Modalities	Remarks
Marketing	CALLAS EU Emotion Stimulus IEMOCAP HUMAINE	According to [225], the advertising campaign focused on emotional content performs twice better than those with rational content. Furthermore, studies have also shown that people who are unable to feel emotion find extremely difficult to make a decision. Therefore, if you can evoke an emotional response in your target audience by studying their facial expressions, voice signals, and body/hand gestures, then there is a high chance that your prospective customer will buy from you. For instance, Nike's "Just do it" fires inner-athlete in everyone and make consumers buy sports goods from them. Similar to Nike, companies such as GE, Cisco, Nike, IBM, AutoDesk, Qualcomm, Facebook, Google, etc., all evoke emotions in their marketing. Therefore, it is suggested to study and analyze databases containing modalities, such as the face, speech, and hand/body gestures, for marketing applications.
Medicine	MAHNOB-HCI emoFBVP DEAP RECOLA	The study of the patient's facial expression, speech tone, body movements, and physiological signals such as ECG, EEG, HR, etc., together can help doctors and nurses discern how the patient is feeling. This can also be useful in the treatment/healing process even if the patient is not physically present. One example of emotion recognition implemented in the field of medicine would be the psychiatrist evaluating the patient's underlying medical conditions by studying their voice and facial expression via smartphones and tablets.
Education and Research	MAHNOB-HCI emoFBVP RECOLA	According to [226,227] the study of emotional states helps in the learning process as they give information about learner's emotion by studying different modalities such as facial expression, speech tone, body/hand/eye movements, and physiological signals such as ECG, EEG, HR, etc. For example, the study of frustration in students related to educational software help improvise the software product. Similarly, the study of boredom in students during their interaction with resources can help ameliorate the study material. Finally, an emotional study on learners also helps in knowing their stereotypes, which further help in adapting the learning paths.
Autonomous Driving Vehicles	MAHNOB-HCI emoFBVP RECOLA NIST IRIS NVIE VIRI KTFE	In the 21st century, with a lot of automotive manufacturers competing to build autonomous driving vehicles, the study of the emotional state of a passenger inside the vehicle during the testing phase would help them build more customer-friendly vehicles. It would also help for future improvements in their vehicle design. For such purpose, the study of facial expression, speech signals, body gestures, and physiological parameters of a passenger inside the car is considered crucial. Moreover, IR databases can also be used for this purpose because it would help know the emotional states during low light conditions or during night time. Furthermore, some of them could also be used in wild conditions such as VIRI.
HCI/ Robotics	MAHNOB-HCI CALLAS EU Emotion Stimulus HUMAINE IEMOCAP	Two most common medium for interaction between human and computer/robot are speech and image. This is achieved through the use of a camera and microphone [228,229]. Therefore, we believe that the study of facial expression, speech signals, eye gaze, and hand/body gestures help in training robots/computer for better interaction with humans.
Aid for Disabled	CALLAS EU Emotion Stimulus HUMAINE IEMOCAP NIST IRIS NVIE VIRI KTFE	Different devices are used to provide aid to a disabled person to make their life easy. It is advised that facial expression, sound signals, and body gestures are important to consider for building aid devices for the disabled because it will help them know the emotional state of the person whom they are talking to and help then react based on that. Furthermore, IR databases can also be used for this application because they are capable of detecting blood flow on the facial regions, even if there is no facial expression on the face. This helps caretakers know very slight or even concealed emotions on the patient with a disability and help them accordingly.
Entertainment	eNTERFACE'05 RML GEMEPFERA 2011 SAVEE AVEC BAUM-1 BAUM-2 VAM SEMAINE SAL NTUA Berlin The USC CreativeIT	People often show their emotions either through face or speech when they watch a movie, sports, news, etc. Therefore, the study of speech signals and facial expressions helps the entertainment industry/ event planner know what people like and dislike, and based on this automatic feedback, they can improvise. Similarly, the study of emotion recognition also helps in movie trailer creation [230]. Additionally, this technique can also be used in gaming. Exergames frameworks and other interactive games have used emotion recognition technology to provide more immense gaming experience [231].
Recommender System	MMI JAFFE CK/CK+ AFEW FER2013 SFEW	The recommender system predicts the preference of the user for any particular product and recommends the best possible option. It makes a recommendation based on the study of facial expressions of people from the previously used/watched contents. For example, during online shopping, change in facial expression for different products can be observed through a camera on laptops, smartphones, and tablets, and the product with positive facial expression can be boosted, while the importance can be reduced for the one with negative facial expression [232]. Apart from the aforementioned field, some of the fields that have successfully applied recommendation system are health care (e.g., emHealth), hospitals, multimedia, web-search, and e-commerce [231].

Table 7. *Cont.*

Application Domain	Databases and Modalities	Remarks
Automatic Surveillance	MMI JAFFE CK/CK+ AFEW FER2013 SFEW CALLAS Expressivity Corpus EU Emotion Stimulus IEMOCAP HUMAINE NIST IRIS VIRI NVIE The KTFE	It is believed that the study of facial features and expression are critical for building of an automatic surveillance system because it helps track the movement of people, identify people at borders, spot criminals, and provide security. Apart from this, databases containing different modalities, such as face, speech, and gestures, can also be studied for automatic surveillance applications at places such as grocery stores, as it can help analyze the age, gender, mood, and behavior of people, which further helps in targeted marketing and product placement. Further, those modalities can also be used in automatic surveillance as seen in movies and videos. Besides the aforementioned databases, IR databases can be a good resource for automatic surveillance during night time.



Facial Speech Eye Movement Body/Hand Gestures Psychological signals.

Author Contributions: Conceptualization: M.F.H.S. and A.Y.J.; Investigation: M.F.H.S.; Methodology: M.F.H.S. and A.Y.J.; Project administration: A.Y.J.; Resources: A.Y.J.; Supervision: A.Y.J. and X.Y.; Validation: M.F.H.S. and P.D.; Writing—original draft: M.F.H.S.; Writing—review and editing, P.D., X.Y. and A.Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the IRB protocol approved by the Institutional Review Board of The University of Toledo (Protocol # 202741-UT, approved on 21 May 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The VIRI dataset is available upon request from the authors. Please visit https://www.yazdan.us/research/repository (accessed on 14 June 2022) for further information.

Acknowledgments: The authors are thankful to Paul A. Hotmer Family CSTAR (Cyber Security and Teaming Research) lab, the Electrical Engineering and Computer Science Department at the University of Toledo and the Computer Science Department at West Texas A&M University.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AEI	Artificial Emotional Intelligence	kNN	k-Nearest Neighbor
AD	Accelerometer Data	LBP-TOP	Local Binary Pattern from Three Orthgonal Planes
AFER	Automatic Facial Expression Recognition	LIF	Local Invariant Feature
ALB	Available but Link Broken	LLD	Low Level Descriptors
ANN	Artificial Neural Network	LSTM	Long Short-Term Memory
AU	Action Units	MCC	Matthew Correlation Coefficient
B-hEOG	Bi-polar Horizontal Electrooculogram	MDR	Multi-Directional Regression
BOS	Blood Oxygen Saturation	MDR	Multi-Directional Regression
BP	Blood Pressure	MEG	Magnetoencephalogram
BS	Brain Signals	MER	Multimodal Emotion Recognition
B-tEMG	Bi-polar trapezius Electromyogram	MFCC	Mel-Frequency Cepstral Coefficients
BVP	Blood Volume Pressure	MKL	Multiple Kernel Learning
CCA	Canonical Correlation Analysis	ML	Machine Learning
CCC	Concordance Correlation Coefficient	MLP	Multi Layer Perceptron
CDBN	Convolutional Deep Belief Network	MSX	Multi Spectral Dynamic Imaging

CNN	Convolutional Neural Network
DB	Database
DBLSTM-RNN	Deep Bi-Directional LSTM-RNN
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DeRL	De-Expression Residue Learning
DL	Deep Learning
DRN	Deep Residual Networks
DT	Dialogu Transcriptions
ECG	Electrocardiogram
ECNN	Ensemble Convolutional Neural Network
EDA	Electro Dermal Activity
EDA	Electro Dermal Activity
EEG	Electroencephalogram
ELM	Extreme Learning Machine
EM	Eye Movements
EMA	Electromagnetic Articulography
EMG	Electromyogram
EOG	Electrooculography
ERC	Emotion Recognition in Conversation
ESD	Emotional Shift Detection
FACS	Facial Action Cading System
FE	Facial Expressions
GECNN	Graph-Embedded Convolutional Neural Network
GSR	Galvanic Skin Response
HCI	Human–Computer Interaction
HMI	Human–Machine Interaction
HMM	Hidden Markov Models
HOG-TOP	Histogram of Oriented Gradients from Three Orthogonal Planes
HR	Heart Rate
IR	Infrared
MuLOT	Multimodal Learning using Optimal Transport
NIR	Near Infrared
OF	Optical Flow
PCA	Principal Component Analysis
PG	Photoplethymogram
PHOG	Pyramid Histograms of Oriented Gradients
PLP	Perceptual Linear Prediction
PPG	Plethysmography
PPS	Psychophysiological signals
PS	Physiological Signals
RASTA	Relative Spectral Features
RBM	Restricted Boltzmann Machines
RBM	Restricted Boltzmann machines
RNN	Recurrent Neural Network
RP	Respiration Pattern
RR	Respiration Rate
RRR	Reduced Rank Regression
RSM	Replicated Softmax Model
SAL	Sensitive Artificial Listener
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SCSR	Skin Conductance Slow Response
SCVCR	Skin Conductance Very Slow Response
SIFT	Scale-Invariant Feature Transform
SKRRR	Sparse Kernel Reduced Rank Regression
SKT	Skin Temperature
SVM	Support Vector Machines
VER	Voice-Based Emotion Recognition
WLD	Weber Local Descriptor
WLD	Weber Local Descriptor
ZCR	Zero Crossing Rate

References

1. Bahreini, K.; Nadolski, R.; Westera, W. Data Fusion for Real-time Multimodal Emotion Recognition through Webcams and Microphones in E-Learning. *Int. J. Hum.–Comput. Interact.* **2016**, *32*, 415–430. [CrossRef]
2. Sun, B.; Li, L.; Zhou, G.; Wu, X.; He, J.; Yu, L.; Li, D.; Wei, Q. Combining multimodal features within a fusion network for emotion recognition in the wild. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 497–502.
3. Xu, C.; Cao, T.; Feng, Z.; Dong, C. Multi-Modal Fusion Emotion Recognition Based on HMM and ANN. *Contemp. Res.-Bus. Technol. Strategy* **2012**, *332*, 541–550.
4. Alonso-Martín, F.; Malfaz, M.; Sequeira, J.; Gorostiza, J.F.; Salichs, M.A. A multimodal emotion detection system during human–robot interaction. *Sensors* **2013**, *13*, 15549–15581. [CrossRef]
5. Kahou, S.E.; Bouthillier, X.; Lamblin, P.; Gulcehre, C.; Michalski, V.; Konda, K.; Jean, S.; Froumenty, P.; Dauphin, Y.; Boulanger-Lewandowski, N.; et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *J. Multimodal User Interfaces* **2016**, *10*, 99–111. [CrossRef]
6. Sun, B.; Li, L.; Zuo, T.; Chen, Y.; Zhou, G.; Wu, X. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; pp. 481–486.
7. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Emotion recognition in the wild with feature fusion and multiple kernel learning. In Proceedings of the 16th International Conference on Multimodal Interaction, Istanbul, Turkey, 12–16 November 2014; pp. 508–513.
8. Tzirakis, P.; Trigeorgis, G.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. End-to-End Multimodal Emotion Recognition using Deep Neural Networks. *arXiv* **2017**, arXiv:1704.08619.
9. Sun, B.; Li, L.; Wu, X.; Zuo, T.; Chen, Y.; Zhou, G.; He, J.; Zhu, X. Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild. *J. Multimodal User Interfaces* **2016**, *10*, 125–137. [CrossRef]

10. Torres, J.M.M.; Stepanov, E.A. Enhanced face/audio emotion recognition: Video and instance level classification using ConvNets and restricted Boltzmann Machines. In Proceedings of the International Conference on Web Intelligence, Leipzig, Germany, 23–26 August 2017; pp. 939–946.
11. Dobrišek, S.; Gajšek, R.; Mihelič, F.; Pavešić, N.; Štruc, V. Towards efficient multi-modal emotion recognition. *Int. J. Adv. Robot. Syst.* **2013**, *10*, 53. [CrossRef]
12. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Fusion of classifier predictions for audio-visual emotion recognition. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 61–66.
13. Kim, Y.; Lee, H.; Provost, E.M. Deep learning for robust feature generation in audiovisual emotion recognition. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 3687–3691.
14. Hossain, M.S.; Muhammad, G. Audio-visual emotion recognition using multi-directional regression and Ridgelet transform. *J. Multimodal User Interfaces* **2016**, *10*, 325–333. [CrossRef]
15. Hossain, M.S.; Muhammad, G.; Alhamid, M.F.; Song, B.; Al-Mutib, K. Audio-visual emotion recognition using big data towards 5G. *Mob. Netw. Appl.* **2016**, *21*, 753–763. [CrossRef]
16. Chen, J.; Chen, Z.; Chi, Z.; Fu, H. Facial expression recognition in video with multiple feature fusion. *IEEE Trans. Affect. Comput.* **2016**, *9*, 38–50. [CrossRef]
17. Yan, J.; Zheng, W.; Xu, Q.; Lu, G.; Li, H.; Wang, B. Sparse Kernel Reduced-Rank Regression for Bimodal Emotion Recognition From Facial Expression and Speech. *IEEE Trans. Multimed.* **2016**, *18*, 1319–1329. [CrossRef]
18. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, New York, NY, USA, 6–9 June 2016; pp. 281–284.
19. Kim, Y. Exploring sources of variation in human behavioral data: Towards automatic audio-visual emotion recognition. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 748–753.
20. Pei, E.; Yang, L.; Jiang, D.; Sahli, H. Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 208–214.
21. Nguyen, D.; Nguyen, K.; Sridharan, S.; Ghasemi, A.; Dean, D.; Fookes, C. Deep spatio-temporal features for multimodal emotion recognition. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 1215–1223.
22. Fu, J.; Mao, Q.; Tu, J.; Zhan, Y. Multimodal shared features learning for emotion recognition by enhanced sparse local discriminative canonical correlation analysis. *Multimed. Syst.* **2017**, *25*, 451–461. [CrossRef]
23. Zhang, S.; Zhang, S.; Huang, T.; Gao, W.; Tian, Q. Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *28*, 3030–3043. [CrossRef]
24. Cid, F.; Manso, L.J.; Núnez, P. A Novel Multimodal Emotion Recognition Approach for Affective Human Robot Interaction. In Proceedings of the Workshop on Multimodal and Semantics for Robotics Systems, Hamburg, Germany, 1 October 2015; pp. 1–9.
25. Haq, S.; Jan, T.; Jehangir, A.; Asif, M.; Ali, A.; Ahmad, N. Bimodal Human Emotion Classification in the Speaker-dependent Scenario. *Pak. Acad. Sci.* **2015**, *52*, 27–38.
26. Gideon, J.; Zhang, B.; Aldeneh, Z.; Kim, Y.; Khorram, S.; Le, D.; Provost, E.M. Wild wild emotion: A multimodal ensemble approach. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, 12–16 November 2016; pp. 501–505.
27. Noroozi, F.; Marjanovic, M.; Njegus, A.; Escalera, S.; Anbarjafari, G. Audio-visual emotion recognition in video clips. *IEEE Trans. Affect. Comput.* **2017**, *10*, 60–75. [CrossRef]
28. Wagner, J.; Andre, E.; Lingenfelser, F.; Kim, J. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Trans. Affect. Comput.* **2011**, *2*, 206–218. [CrossRef]
29. Ghayoumi, M.; Bansal, A.K. Multimodal architecture for emotion in robots using deep learning. In Proceedings of the Future Technologies Conference (FTC), San Francisco, CA, USA, 6–7 December 2016; pp. 901–907.
30. Kessous, L.; Castellano, G.; Caridakis, G. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *J. Multimodal User Interfaces* **2010**, *3*, 33–48. [CrossRef]
31. Yoshitomi, Y.; Kim, S.I.; Kawano, T.; Kilazoe, T. Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face. In Proceedings of the Proceedings 9th IEEE International Workshop on Robot and Human Interactive Communication. IEEE RO-MAN 2000 (Cat. No.00TH8499), Osaka, Japan, 27–29 September 2000; pp. 178–183.
32. Kitazoe, T.; Kim, S.I.; Yoshitomi, Y.; Ikeda, T. Recognition of emotional states using voice, face image and thermal image of face. In Proceedings of the Sixth International Conference on Spoken Language Processing, Beijing, China, 16–20 October 2000.
33. Shah, M.; Chakrabarti, C.; Spanias, A. A multi-modal approach to emotion recognition using undirected topic models. In Proceedings of the 2014 IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne, Australia, 1–5 June 2014; pp. 754–757.

34. Verma, G.K.; Tiwary, U.S. Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals. *NeuroImage* **2014**, *102*, 162–172. [CrossRef]
35. Keren, G.; Kirschstein, T.; Marchi, E.; Ringeval, F.; Schuller, B. End-to-end learning for dimensional emotion recognition from physiological signals. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 985–990.
36. Yin, Z.; Zhao, M.; Wang, Y.; Yang, J.; Zhang, J. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Comput. Methods Programs Biomed.* **2017**, *140*, 93–110. [CrossRef]
37. Dai, Y.; Wang, X.; Zhang, P.; Zhang, W. Wearable Biosensor Network Enabled Multimodal Daily-life Emotion Recognition Employing Reputation-driven Imbalanced Fuzzy Classification. *Measurement* **2017**, *109*, 408–424. [CrossRef]
38. Kortelainen, J.; Tiinanen, S.; Huang, X.; Li, X.; Laukka, S.; Pietikäinen, M.; Seppänen, T. Multimodal emotion recognition by combining physiological signals and facial expressions: A preliminary study. In Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, San Diego, CA, USA, 28 August–1 September 2012; pp. 5238–5241.
39. Hess, U.; Thibault, P. Darwin and emotion expression. *Am. Psychol.* **2009**, *64*, 120. [CrossRef]
40. Laird, J.D.; Lacasse, K. Bodily influences on emotional feelings: Accumulating evidence and extensions of William James's theory of emotion. *Emot. Rev.* **2014**, *6*, 27–34. [CrossRef]
41. Corneanu, C.A.; Simón, M.O.; Cohn, J.F.; Guerrero, S.E. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1548–1568. [CrossRef]
42. Ekman, P. Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique. *Psychol. Bull.* **1994**, *115*, 268–287. [CrossRef]
43. Ekman, P.; Friesen, W.V.; Hager, J. *Investigator's Guide to the Facial Action Coding System*; Consulting Psychologists Press: Palo Alto, CA, USA, 1978.
44. Mase, K. Recognition of facial expression from optical flow. *IEICE Trans. Inf. Syst.* **1991**, *74*, 3474–3483.
45. Lanitis, A.; Taylor, C.J.; Cootes, T.F. A unified approach to coding and interpreting face images. In Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 20–23 June 1995; pp. 368–373.
46. Black, M.J.; Yacoob, Y. Tracking and recognizing rigid and non-rigid facial motions using local parametric models of image motion. In Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 20–23 June 1995; pp. 374–381.
47. Rosenblum, M.; Yacoob, Y.; Davis, L.S. Human expression recognition from motion using a radial basis function network architecture. *IEEE Trans. Neural Netw.* **1996**, *7*, 1121–1138. [CrossRef]
48. Essa, I.A.; Pentland, A.P. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 757–763. [CrossRef]
49. Yacoob, Y.; Davis, L.S. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.* **1996**, *18*, 636–642. [CrossRef]
50. Lang, P.J.; Bradley, M.M.; Cuthbert, B.N. International affective picture system (IAPS): Technical manual and affective ratings. *NIMH Cent. Study Emot. Atten.* **1997**, *1*, 39–58.
51. Santamaria-Granados, L.; Munoz-Organero, M.; Ramirez-Gonzalez, G.; Abdulhay, E.; Arunkumar, N. Using deep convolutional neural network for emotion detection on a physiological signals dataset (AMIGOS). *IEEE Access* **2018**, *7*, 57–67. [CrossRef]
52. Sourina, O.; Liu, Y. A fractal-based algorithm of emotion recognition from EEG using arousal-valence model. In Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing, Rome, Italy, 26–29 January 2011; Volume 2, pp. 209–214.
53. Liu, Y.; Sourina, O.; Nguyen, M.K. Real-time EEG-Based emotion recognition and its applications. In *Transactions on Computational Science XII*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 256–277.
54. Alhagry, S.; Fahmy, A.A.; El-Khoribi, R.A. Emotion recognition based on EEG using LSTM recurrent neural network. *Emotion* **2017**, *8*, 355–358. [CrossRef]
55. Zhang, Q.; Chen, X.; Zhan, Q.; Yang, T.; Xia, S. Respiration-based emotion recognition with deep learning. *Comput. Ind.* **2017**, *92*, 84–90. [CrossRef]
56. Aleksic, P.S.; Katsaggelos, A.K. Automatic facial expression recognition using facial animation parameters and multistream HMMs. *IEEE Trans. Inf. Forensics Secur.* **2006**, *1*, 3–11. [CrossRef]
57. Tian, Y.I.; Kanade, T.; Cohn, J.F. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 97–115. [CrossRef] [PubMed]
58. Matsugu, M.; Mori, K.; Mitari, Y.; Kaneda, Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Netw.* **2003**, *16*, 555–559. [CrossRef]
59. Yin, L.; Wei, X.; Sun, Y.; Wang, J.; Rosato, M.J. A 3D facial expression database for facial behavior research. In Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), Southampton, UK, 10–12 April 2006; pp. 211–216. Available online: https://dl.acm.org/doi/abs/10.5555/1126250.1126340 (accessed on 14 June 2022).
60. Mandal, T.; Majumdar, A.; Wu, Q.J. Face recognition by curvelet based feature extraction. In Proceedings of the International Conference Image Analysis and Recognition, Montreal, QC, Canada, 22–24 August 2007; pp. 806–817.

61. Li, C.; Soares, A. Automatic facial expression recognition using 3D faces. *Int. J. Eng. Res. Innov.* **2011**, *3*, 30–34.
62. Jain, D.K.; Shamsolmoali, P.; Sehdev, P. Extended deep neural network for facial emotion recognition. *Pattern Recognit. Lett.* **2019**, *120*, 69–74. [CrossRef]
63. Chen, L.; Zhou, M.; Su, W.; Wu, M.; She, J.; Hirota, K. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. *Inf. Sci.* **2018**, *428*, 49–61. [CrossRef]
64. Bazrafkan, S.; Nedelcu, T.; Filipczuk, P.; Corcoran, P. Deep learning for facial expression recognition: A step closer to a smartphone that knows your moods. In Proceedings of the 2017 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 8–11 January 2017; pp. 217–220.
65. Zhang, T.; Zheng, W.; Cui, Z.; Zong, Y.; Yan, J.; Yan, K. A deep neural network-driven feature learning method for multi-view facial expression recognition. *IEEE Trans. Multimed.* **2016**, *18*, 2528–2536. [CrossRef]
66. Sebe, N.; Cohen, I.; Gevers, T.; Huang, T.S. Multimodal approaches for emotion recognition: A survey. In Proceedings of the SPIE Internet Imaging VI, San Jose, CA, USA, 16–20 January 2005; Volume 5670, pp. 56–67.
67. Busso, C.; Mariooryad, S.; Metallinou, A.; Narayanan, S. Iterative feature normalization scheme for automatic emotion detection from speech. *IEEE Trans. Affect. Comput.* **2013**, *4*, 386–397. [CrossRef]
68. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Trans. Multimed.* **2014**, *16*, 2203–2213. [CrossRef]
69. Gangeh, M.J.; Fewzee, P.; Ghodsi, A.; Kamel, M.S.; Karray, F. Multiview supervised dictionary learning in speech emotion recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1056–1068. [CrossRef]
70. Wang, K.; An, N.; Li, B.N.; Zhang, Y.; Li, L. Speech emotion recognition using Fourier parameters. *IEEE Trans. Affect. Comput.* **2015**, *6*, 69–75. [CrossRef]
71. Fayek, H.M.; Lech, M.; Cavedon, L. Towards real-time speech emotion recognition using deep neural networks. In Proceedings of the 2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS), Cairns, Australia, 14–16 December 2015; pp. 1–5.
72. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 1089–1093.
73. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.
74. Rozgić, V.; Vitaladevuni, S.N.; Prasad, R. Robust EEG emotion classification using segment level decision fusion. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 1286–1290.
75. Lakens, D. Using a smartphone to measure heart rate changes during relived happiness and anger. *IEEE Trans. Affect. Comput.* **2013**, *4*, 238–241. [CrossRef]
76. Hernandez, J.; McDuff, D.; Fletcher, R.; Picard, R.W. Inside-out: Reflecting on your inner state. In Proceedings of the 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), San Diego, CA, USA, 18–22 March 2013; pp. 324–327.
77. Fridlund, A.; Izard, C.E. Electromyographic studies of facial expressions of emotions and patterns of emotions. In *Social Psychophysiology: A Sourcebook*; Guilford Press: New York, NY, USA, 1983; pp. 243–286.
78. Lin, W.; Li, C.; Sun, S. Deep convolutional neural network for emotion recognition using EEG and peripheral physiological signal. In Proceedings of the International Conference on Image and Graphics, Shanghai, China, 13–15 September 2017; pp. 385–394.
79. Paleari, M.; Chellali, R.; Huet, B. Features for multimodal emotion recognition: An extensive study. In Proceedings of the 2010 IEEE Conference on Cybernetics and Intelligent Systems, Singapore, 28–30 June 2010; pp. 90–95.
80. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, p. I.
81. De Silva, L.C.; Ng, P.C. Bimodal emotion recognition. In Proceedings of the Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), Grenoble, France, 26–30 March 2000; pp. 332–335.
82. Chen, L.S.; Huang, T.S. Emotional expressions in audiovisual human computer interaction. In Proceedings of the 2000 IEEE International Conference on Multimedia and Expo, ICME2000, Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532), New York, NY, USA, 30 July–2 August 2000; Volume 1, pp. 423–426.
83. Caridakis, G.; Castellano, G.; Kessous, L.; Raouzaiou, A.; Malatesta, L.; Asteriadis, S.; Karpouzis, K. Multimodal emotion recognition from expressive faces, body gestures and speech. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*; Springer: Boston, MA, USA, 2007; pp. 375–388.
84. Tang, K.; Tie, Y.; Yang, T.; Guan, L. Multimodal emotion recognition (MER) system. In Proceedings of the 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), Toronto, ON, Canada, 4–7 May 2014; pp. 1–6.
85. Soleymani, M.; Lichtenauer, J.; Pun, T.; Pantic, M. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.* **2012**, *3*, 42–55. [CrossRef]

86. Ranganathan, H.; Chakraborty, S.; Panchanathan, S. Multimodal emotion recognition using deep learning architectures. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–10 Macrh 2016; pp. 1–9.
87. Lee, H.; Grosse, R.; Ranganath, R.; Ng, A.Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 14–18 June 2009; pp. 609–616.
88. Poria, S.; Chaturvedi, I.; Cambria, E.; Hussain, A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In Proceedings of the 2016 IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 12–15 December 2016; pp. 439–448.
89. Dataset 02: IRIS Thermal/Visible Face Database 2016. Available online: http://vcipl-okstate.org/pbvs/bench/ (accessed on 14 June 2022).
90. Dataset 01: NIST Thermal/Visible Face Database 2012. Available online: https://www.google.com.hk/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjRhJn--LP4AhVFCd4KHYiOAhgQFnoECAYQAQ&url=https%3A%2F%2Fwww.nist.gov%2Fdocument%2Fklare-nistdatasets2015pdf&usg=AOvVaw0O-vRUczPwxCTSp2_SWWe7 (accessed on 14 June 2022).
91. Wang, S.; Liu, Z.; Lv, S.; Lv, Y.; Wu, G.; Peng, P.; Chen, F.; Wang, X. A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Trans. Multimed.* **2010**, *12*, 682–691. [CrossRef]
92. Nguyen, H.; Kotani, K.; Chen, F.; Le, B. A thermal facial emotion database and its analysis. In Proceedings of the Pacific-Rim Symposium on Image and Video Technology, Guanajuato, México, 28 October–1 November 2013; pp. 397–408.
93. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335. [CrossRef]
94. Correa, J.A.M.; Abadi, M.K.; Sebe, N.; Patras, I. Amigos: A dataset for affect, personality and mood research on individuals and groups. *IEEE Trans. Affect. Comput.* **2018**, *12*, 479–493. [CrossRef]
95. EMOTIV | Brain Data Measuring Hardware and Software Solutions. Available online: https://www.emotiv.com/ (accessed on 18 May 2020).
96. SHIMMER | Wearable Sensor Technology | Wireless IMU | ECG | EMG | GSR. Available online: http://www.shimmersensing.com/ (accessed on 18 May 2020).
97. Subramanian, R.; Wache, J.; Abadi, M.K.; Vieriu, R.L.; Winkler, S.; Sebe, N. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Trans. Affect. Comput.* **2016**, *9*, 147–160. [CrossRef]
98. Caridakis, G.; Wagner, J.; Raouzaiou, A.; Curto, Z.; Andre, E.; Karpouzis, K. A multimodal corpus for gesture expressivity analysis. In Proceedings of the Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, LREC, Valetta, Malta, 18 May 2010.
99. Caridakis, G.; Wagner, J.; Raouzaiou, A.; Lingenfelser, F.; Karpouzis, K.; Andre, E. A cross-cultural, multimodal, affective corpus for gesture expressivity analysis. *J. Multimodal User Interfaces* **2013**, *7*, 121–134. [CrossRef]
100. Markova, V.; Ganchev, T.; Kalinkov, K. CLAS: A Database for Cognitive Load, Affect and Stress Recognition. In Proceedings of the 2019 International Conference on Biomedical Innovations and Applications (BIA), Varna, Bulgaria, 8–9 November 2019; pp. 1–4.
101. SHIMMER3 ECG Unit | Wearable ECG Sensor | Wireless ECG Sensor | Electrocardiogram. Available online: https://www.shimmersensing.com/products/shimmer3-ecg-sensor (accessed on 19 May 2020).
102. Shimmer3 GSR+ Sensor. Available online: http://www.shimmersensing.com/shimmer3-gsr-sensor/ (accessed on 19 May 2020).
103. Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
104. Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; Patras, I. Deap: A database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* **2012**, *3*, 18–31. [CrossRef]
105. Abadi, M.K.; Subramanian, R.; Kia, S.M.; Avesani, P.; Patras, I.; Sebe, N. DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Trans. Affect. Comput.* **2015**, *6*, 209–222. [CrossRef]
106. O'Reilly, H.; Pigat, D.; Fridenson, S.; Berggren, S.; Tal, S.; Golan, O.; Bölte, S.; Baron-Cohen, S.; Lundqvist, D. The EU-emotion stimulus set: A validation study. *Behav. Res. Methods* **2016**, *48*, 567–576. [CrossRef] [PubMed]
107. Chen, J.; Wang, C.; Wang, K.; Yin, C.; Zhao, C.; Xu, T.; Zhang, X.; Huang, Z.; Liu, M.; Yang, T. HEU Emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Comput. Appl.* **2021**, *33*, 8669–8685. [CrossRef]
108. Huang, X.; Kortelainen, J.; Zhao, G.; Li, X.; Moilanen, A.; Seppänen, T.; Pietikäinen, M. Multi-modal emotion analysis from facial expressions and electroencephalogram. *Comput. Vis. Image Underst.* **2016**, *147*, 114–124. [CrossRef]
109. Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv* **2018**, arXiv:1810.02508.
110. Chen, S.Y.; Hsu, C.C.; Kuo, C.C.; Ku, L.W. Emotionlines: An emotion corpus of multi-party conversations. *arXiv* **2018**, arXiv:1802.08379.
111. Tu, G.; Wen, J.; Liu, C.; Jiang, D.; Cambria, E. Context-and sentiment-aware networks for emotion recognition in conversation. *IEEE Trans. Artif. Intell.* **2022**. [CrossRef]

112. Zhang, Z.; Girard, J.M.; Wu, Y.; Zhang, X.; Liu, P.; Ciftci, U.; Canavan, S.; Reale, M.; Horowitz, A.; Yang, H.; et al. Multimodal spontaneous emotion corpus for human behavior analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3438–3446.
113. Yang, H.; Ciftci, U.; Yin, L. Facial expression recognition by de-expression residue learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2168–2177.
114. Jannat, R.; Tynes, I.; Lime, L.L.; Adorno, J.; Canavan, S. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore, 8–12 October 2018; pp. 956–959.
115. Song, T.; Zheng, W.; Lu, C.; Zong, Y.; Zhang, X.; Cui, Z. MPED: A multi-modal physiological emotion database for discrete emotion recognition. *IEEE Access* **2019**, *7*, 12177–12191. [CrossRef]
116. Song, T.; Zheng, W.; Liu, S.; Zong, Y.; Cui, Z.; Li, Y. Graph-Embedded Convolutional Neural Network for Image-based EEG Emotion Recognition. *IEEE Trans. Emerg. Top. Comput.* **2021**. [CrossRef]
117. Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; Poria, S. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv* **2019**, arXiv:1906.01815.
118. Sarcasm | Psychology Today. Available online: https://www.psychologytoday.com/us/blog/stronger-the-broken-places/201907/sarcasm (accessed on 17 May 2020).
119. Zhang, Y.; Tiwari, P.; Rong, L.; Chen, R.; AlNajem, N.A.; Hossain, M.S. Affective Interaction: Attentive Representation Learning for Multi-Modal Sentiment Classification. In *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*; ACM: New York, NY, USA, 2021 .
120. Pramanick, S.; Roy, A.; Patel, V.M. Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikola, HI, USA, 5–7 January 2022; pp. 3930–3940.
121. Chou, H.C.; Lin, W.C.; Chang, L.C.; Li, C.C.; Ma, H.P.; Lee, C.C. NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus. In Proceedings of the 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, 23–26 October 2017; pp. 292–298.
122. Hsu, J.H.; Su, M.H.; Wu, C.H.; Chen, Y.H. Speech emotion recognition considering nonverbal vocalization in affective conversations. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1675–1686. [CrossRef]
123. Perepelkina, O.; Kazimirova, E.; Konstantinova, M. RAMAS: Russian multimodal corpus of dyadic interaction for affective computing. In Proceedings of the International Conference on Speech and Computer, Leipzig, Germany, 18–22 September 2018; pp. 501–510.
124. Sloetjes, H.; Wittenburg, P. Annotation by category-ELAN and ISO DCR. In Proceedings of the 6th international Conference on Language Resources and Evaluation (LREC 2008), Marrakech, Morocco, 28–30 May 2008.
125. Verkholyak, O.; Dvoynikova, A.; Karpov, A. A Bimodal Approach for Speech Emotion Recognition using Audio and Text. *J. Internet Serv. Inf. Secur.* **2021**, *11*, 80–96.
126. Ringeval, F.; Sonderegger, A.; Sauer, J.; Lalanne, D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, China, 22–26 April 2013; pp. 1–8.
127. Mencattini, A.; Ringeval, F.; Schuller, B.; Martinelli, E.; Di Natale, C. Continuous monitoring of emotions by a multimodal cooperative sensor system. *Procedia Eng.* **2015**, *120*, 556–559. [CrossRef]
128. Ganchev, T.; Markova, V.; Lefterov, I.; Kalinin, Y. Overall Design of the SLADE Data Acquisition System. In Proceedings of the International Conference on Intelligent Information Technologies for Industry, Sirius, Russia, 30September–4 October 2017; pp. 56–65.
129. Valstar, M.; Schuller, B.; Smith, K.; Eyben, F.; Jiang, B.; Bilakhia, S.; Schnieder, S.; Cowie, R.; Pantic, M. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, Barcelona, Spain, 21 October 2013; pp. 3–10.
130. Valstar, M.; Schuller, B.; Smith, K.; Almaev, T.; Eyben, F.; Krajewski, J.; Cowie, R.; Pantic, M. Avec 2014: 3d dimensional affect and depression recognition challenge. In Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Orlando, FL, USA, 7 November 2014; pp. 3–10.
131. Tian, L.; Moore, J.; Lai, C. Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features. In Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, USA, 13–16 December 2016; pp. 565–572.
132. Zhalehpour, S.; Onder, O.; Akhtar, Z.; Erdem, C.E. BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States. *IEEE Trans. Affect. Comput.* **2017**, *8*, 300–313. [CrossRef]
133. Zhang, L.; Walter, S.; Ma, X.; Werner, P.; Al-Hamadi, A.; Traue, H.C.; Gruss, S. "BioVid Emo DB": A multimodal database for emotion analyses validated by subjective ratings. In Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 6–9 December 2016; pp. 1–6.
134. Prabha, R.; Anandan, P.; Sivarajeswari, S.; Saravanakumar, C.; Babu, D.V. Design of an Automated Recurrent Neural Network for Emotional Intelligence Using Deep Neural Networks. In Proceedings of the 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 January 2022; pp. 1061–1067.

135. Li, Y.; Tao, J.; Chao, L.; Bao, W.; Liu, Y. CHEAVD: A Chinese natural emotional audio–visual database. *J. Ambient. Intell. Humaniz. Comput.* **2017**, *8*, 913–924. [CrossRef]
136. Li, Y.; Tao, J.; Schuller, B.; Shan, S.; Jiang, D.; Jia, J. Mec 2017: Multimodal emotion recognition challenge. In Proceedings of the 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), Beijing, China, 20–22 May 2018; pp. 1–5.
137. Wang, C.; Ren, Y.; Zhang, N.; Cui, F.; Luo, S. Speech emotion recognition based on multi-feature and multi-lingual fusion. *Multimed. Tools Appl.* **2022**, *81*, 4897–4907. [CrossRef]
138. Liang, J.; Chen, S.; Zhao, J.; Jin, Q.; Liu, H.; Lu, L. Cross-culture multimodal emotion recognition with adversarial learning. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 4000–4004.
139. Katsigiannis, S.; Ramzan, N. DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. *IEEE J. Biomed. Health Inform.* **2017**, *22*, 98–107. [CrossRef]
140. Badcock, N.A.; Mousikou, P.; Mahajan, Y.; De Lissa, P.; Thie, J.; McArthur, G. Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. *PeerJ* **2013**, *1*, e38. [CrossRef]
141. Ekanayake, H. P300 and Emotiv EPOC: Does Emotiv EPOC Capture Real EEG? 2010. Available online: http://neurofeedback.visaduma.info/emotivresearch.htm (accessed on 6 June 2022).
142. Burns, A.; Greene, B.R.; McGrath, M.J.; O'Shea, T.J.; Kuris, B.; Ayer, S.M.; Stroiescu, F.; Cionca, V. SHIMMER™–A wireless sensor platform for noninvasive biomedical research. *IEEE Sens. J.* **2010**, *10*, 1527–1534. [CrossRef]
143. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The enterface'05 audio-visual emotion database. In Proceedings of the 22nd International Conference on Data Engineering Workshops, Washington, DC, USA, 3–7 April 2006; p. 8.
144. Gunes, H.; Piccardi, M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–26 August 2006; Volume 1, pp. 1148–1153.
145. Karatay, B.; Bestepe, D.; Sailunaz, K.; Ozyer, T.; Alhajj, R. A Multi-Modal Emotion Recognition System Based on CNN-Transformer Deep Learning Technique. In Proceedings of the 2022 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 1–3 March 2022; pp. 145–150.
146. Lucey, P.; Cohn, J.F.; Kanade, T.; Saragih, J.; Ambadar, Z.; Matthews, I. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), San Francisco, CA, USA, 13–18 June 2010; pp. 94–101.
147. Valstar, M.F.; Jiang, B.; Mehu, M.; Pantic, M.; Scherer, K. The first facial expression recognition and analysis challenge. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011), Santa Barbara, CA, USA, 21–23 March 2011; pp. 921–926.
148. Bänziger, T.; Scherer, K.R. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Bluepr. Affect. Comput. Sourceb.* **2010**, *2010*, 271–294.
149. Douglas-Cowie, E.; Cowie, R.; Sneddon, I.; Cox, C.; Lowry, O.; Mcrorie, M.; Martin, J.C.; Devillers, L.; Abrilian, S.; Batliner, A.; et al. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. *Affect. Comput. Intell. Interact.* **2007**, *4738*, 488–500.
150. Baveye, Y.; Bettinelli, J.N.; Dellandréa, E.; Chen, L.; Chamaret, C. A large video database for computational models of induced emotion. In Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 2–5 September 2013; pp. 13–18.
151. Baveye, Y.; Dellandrea, E.; Chamaret, C.; Chen, L. Liris-accede: A video database for affective content analysis. *IEEE Trans. Affect. Comput.* **2015**, *6*, 43–55. [CrossRef]
152. Livingstone, S.R.; Russo, F.A. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* **2018**, *13*, e0196391. [CrossRef]
153. Iqbal, A.; Barua, K. A Real-time Emotion Recognition from Speech using Gradient Boosting. In Proceedings of the 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox'sBazar, Bangladesh, 7–9 February 2019; pp. 1–5.
154. Haque, A.; Guo, M.; Verma, P.; Fei-Fei, L. Audio-linguistic embeddings for spoken sentences. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 7355–7359.
155. Wang, Y.; Guan, L. Recognizing human emotional state from audiovisual signals. *IEEE Trans. Multimed.* **2008**, *10*, 936–946. [CrossRef]
156. Gievska, S.; Koroveshovski, K.; Tagasovska, N. Bimodal feature-based fusion for real-time emotion recognition in a mobile context. In Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), Xi'an, China, 21–24 September 2015; pp. 401–407.
157. Gunes, H.; Pantic, M. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In Proceedings of the Intelligent Virtual Agents, Philadelphia, PA, USA, 20–22 September 2010; pp. 371–377.
158. Haq, S.; Jackson, P.J. Multimodal emotion recognition. In *Machine Audition: Principles, Algorithms and Systems*; IGI Global: Hershey, PA, USA, 2010; pp. 398–423.

159. Zheng, W.L.; Lu, B.L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* **2015**, *7*, 162–175. [CrossRef]
160. Liu, W.; Qiu, J.L.; Zheng, W.L.; Lu, B.L. Multimodal Emotion Recognition Using Deep Canonical Correlation Analysis. *arXiv* **2019**, arXiv:1908.05349.
161. Duan, R.N.; Zhu, J.Y.; Lu, B.L. Differential entropy feature for EEG-based emotion classification. In Proceedings of the 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), San Diego, CA, USA, 6–8 November 2013; pp. 81–84.
162. Zheng, W.L.; Liu, W.; Lu, Y.; Lu, B.L.; Cichocki, A. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* **2018**, *49*, 1110–1122. [CrossRef]
163. Li, T.H.; Liu, W.; Zheng, W.L.; Lu, B.L. Classification of five emotions from EEG and eye movement signals: Discrimination ability and stability over time. In Proceedings of the 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER), San Francisco, CA, USA, 20–23 March 2019; pp. 607–610.
164. Zheng, W.L.; Lu, B.L. A multimodal approach to estimating vigilance using EEG and forehead EOG. *J. Neural Eng.* **2017**, *14*, 026017. [CrossRef]
165. McKeown, G.; Valstar, M.; Cowie, R.; Pantic, M.; Schroder, M. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Trans. Affect. Comput.* **2012**, *3*, 5–17. [CrossRef]
166. Metallinou, A.; Yang, Z.; Lee, C.c.; Busso, C.; Carnicke, S.; Narayanan, S. The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Lang. Resour. Eval.* **2016**, *50*, 497–521. [CrossRef]
167. Chang, C.M.; Lee, C.C. Fusion of multiple emotion perspectives: Improving affect recognition through integrating cross-lingual emotion information. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5820–5824.
168. Grimm, M.; Kroschel, K.; Narayanan, S. The Vera am Mittag German audio-visual emotional speech database. In Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, Hannover, Germany, 23–26 June 2008; pp. 865–868.
169. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimed.* **2012**, *19*, 34–41. [CrossRef]
170. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.F.; Weiss, B. A database of german emotional speech. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; Volume 5, pp. 1517–1520.
171. Staroniewicz, P.; Majewski, W. Polish emotional speech database–recording and preliminary validation. In *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 42–49.
172. Lee, S.; Yildirim, S.; Kazemzadeh, A.; Narayanan, S. An articulatory study of emotional speech production. In Proceedings of the Interspeech, Lisbon, Portugal, 4–8 September 2005; pp. 497–500.
173. Strapparava, C.; Mihalcea, R. Semeval-2007 task 14: Affective text. In Proceedings of the 4th International Workshop on Semantic Evaluations. Association for Computational Linguistics, Prague, Czech Republic, 23–24 June 2007; pp. 70–74.
174. Wallbott, H.G.; Scherer, K.R. How universal and specific is emotional experience? Evidence from 27 countries on five continents. *Soc. Sci. Inf.* **1986**, *25*, 763–795. [CrossRef]
175. Kanade, T.; Cohn, J.F.; Tian, Y. Comprehensive database for facial expression analysis. In Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition Grenoble, France, 26–30 March 2000; pp. 46–53.
176. Ekman, P.; Friesen, W.V. Facial Action Coding System 1977. Available online: https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft27734-000 (accessed on 14 June 2022) [CrossRef]
177. Ekman, P.; Friesen, W.V.; Hager, J.C. FACS Investigator's Guide. 2002, 96 Chapter 4 pp 29. Available online: https://www.scirp.org/%28S%28i43dyn45teexjx455qlt3d2q%29%29/reference/ReferencesPapers.aspx?ReferenceID=1850657 (accessed on 14 June 2022).
178. Ranganathan, H.; Chakraborty, S.; Panchanathan, S. Transfer of multimodal emotion features in deep belief networks. In Proceedings of the 2016 50th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 6–9 November 2016; pp. 449–453.
179. Wen, G.; Hou, Z.; Li, H.; Li, D.; Jiang, L.; Xun, E. Ensemble of Deep Neural Networks with Probability-Based Fusion for Facial Expression Recognition. *Cogn. Comput.* **2017**, *9*, 597–610. [CrossRef]
180. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Korea, 3–7 November 2013; pp. 117–124.
181. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 443–449.
182. Dailey, M.N.; Joyce, C.; Lyons, M.J.; Kamachi, M.; Ishi, H.; Gyoba, J.; Cottrell, G.W. Evidence and a computational explanation of cultural differences in facial expression recognition. *Emotion* **2010**, *10*, 874. [CrossRef]
183. Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 14–16 April 1998; pp. 200–205.
184. Lyons, M.J.; Budynek, J.; Akamatsu, S. Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 1357–1362. [CrossRef]

185. Pantic, M.; Valstar, M.; Rademaker, R.; Maat, L. Web-based database for facial expression analysis. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; p. 5.
186. Valstar, M.; Pantic, M. Induced disgust, happiness and surprise: An addition to the mmi facial expression database. In Proceedings of the 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Valletta, Malta, 23 May 2010; p. 65.
187. Dhall, A.; Goecke, R.; Lucey, S.; Gedeon, T. Static Facial Expressions In The Wild: Data and Experiment Protocol. Available online: http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.671.1708 (accessed on 4 June 2022).
188. Yin, G.; Sun, S.; Yu, D.; Li, D.; Zhang, K. A Multimodal Framework for Large-Scale Emotion Recognition by Fusing Music and Electrodermal Activity Signals. *ACM Trans. Multimed. Comput. Commun. Appl. (Tomm)* **2022**, *18*, 1–23. [CrossRef]
189. Udovičić, G.; Đerek, J.; Russo, M.; Sikora, M. Wearable emotion recognition system based on GSR and PPG signals. In Proceedings of the 2nd International Workshop on Multimedia for Personal Health and Health Care, Mountain View, CA, USA, 23 October 2017; pp. 53–59.
190. Radhika, K.; Oruganti, V.R.M. Deep Multimodal Fusion for Subject-Independent Stress Detection. In Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 28–29 January 2021; pp. 105–109.
191. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 1359–1367.
192. Pham, M.; Do, H.M.; Su, Z.; Bishop, A.; Sheng, W. Negative emotion management using a smart shirt and a robot assistant. *IEEE Robot. Autom. Lett.* **2021**, *6*, 4040–4047. [CrossRef]
193. Sun, B.; Cao, S.; Li, L.; He, J.; Yu, L. Exploring multimodal visual features for continuous affect recognition. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, Amsterdam, The Netherlands, 16 October 2016; pp. 83–88.
194. Erdem, C.E.; Turan, C.; Aydin, Z. BAUM-2: A multilingual audio-visual affective face database. *Multimed. Tools Appl.* **2015**, *74*, 7429–7459. [CrossRef]
195. Dar, M.N.; Akram, M.U.; Khawaja, S.G.; Pujari, A.N. CNN and LSTM-based emotion charting using physiological signals. *Sensors* **2020**, *20*, 4551. [CrossRef]
196. Siddharth, S.; Jung, T.P.; Sejnowski, T.J. Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *IEEE Trans. Affect. Comput.* **2019**, *13*, 96–107. [CrossRef]
197. Yi, Y.; Wang, H.; Tang, P. Unified Multi-Stage Fusion Network for Affective Video Content Analysis. Available at SSRN 4080629. Available online: https://ssrn.com/abstract=4080629 (accessed on 14 June 2022).
198. McKeown, G.; Valstar, M.F.; Cowie, R.; Pantic, M. The SEMAINE corpus of emotionally coloured character interactions. In Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME), Singapore, 19–23 July 2010; pp. 1079–1084.
199. Siddiqui, M.F.H.; Javaid, A.Y. A Multimodal Facial Emotion Recognition Framework through the Fusion of Speech with Visible and Infrared Images. *Multimodal Technol. Interact.* **2020**, *4*, 46. [CrossRef]
200. Andrew Ng: Why AI Is the New Electricity | The Dish. Available online: https://news.stanford.edu/thedish/2017/03/14/andrew-ng-why-ai-is-the-new-electricity/ (accessed on 3 May 2018).
201. Emotional Intelligence is the Future of Artificial Intelligence: Fjord | ZDNet. Available online: http://www.zdnet.com/article/emotional-intelligence-is-the-future-of-artificial-intelligence-fjord/ (accessed on 3 May 2018).
202. Synced | Emotional Intelligence is the Future of Artificial Intelligence. Available online: https://syncedreview.com/2017/03/14/emotional-intelligence-is-the-future-of-artificial-intelligence/ (accessed on 3 May 2018).
203. Olszewska, J.I. Automated Face Recognition: Challenges and Solutions. In *Pattern Recognition-Analysis and Applications*; InTech: London, UK, 2016.
204. Lie to Me | Paul Ekman Group. Available online: https://www.paulekman.com/lie-to-me/ (accessed on 3 June 2018).
205. Arellano, D.; Varona, J.; Perales, F.J. Emotional Context? Or Contextual Emotions? In *Handbook of Research on Synthesizing Human Emotion in Intelligent Systems and Robotics*; IGI Global: Hershey, PA, USA, 2015; pp. 366–385.
206. Bullington, J. 'Affective'computing and emotion recognition systems: The future of biometric surveillance? In Proceedings of the 2nd Annual Conference on Information Security Curriculum Development, Kennesaw, GA, USA, 23–24 September 2005; pp. 95–99.
207. Disney Is Using Facial Recognition to Predict How You'll React to Movies. Available online: https://mashable.com/2017/07/27/disney-facial-recognition-prediction-movies/#aoVIBBcxxmqI (accessed on 3 June 2018).
208. Xie, Z.; Guan, L. Multimodal information fusion of audiovisual emotion recognition using novel information theoretic tools. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
209. Wu, C.H.; Lin, J.C.; Wei, W.L. Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies. *APSIPA Trans. Signal Inf. Process.* **2014**, *3*, e12. [CrossRef]
210. Wang, Y.; Guan, L.; Venetsanopoulos, A.N. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Trans. Multimed.* **2012**, *14*, 597–607. [CrossRef]

211. Mehmood, R.M.; Lee, H.J. A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns. *Comput. Electr. Eng.* **2016**, *53*, 444–457. [CrossRef]
212. Pramerdorfer, C.; Kampel, M. Facial Expression Recognition using Convolutional Neural Networks: State of the Art. *arXiv* **2016**, arXiv:1612.02903.
213. Lang, P.; Bradley, M.M. The International Affective Picture System (IAPS) in the study of emotion and attention. *Handb. Emot. Elicitation Assess.* **2007**, *29*, 70–73.
214. Kim, B.K.; Dong, S.Y.; Roh, J.; Kim, G.; Lee, S.Y. Fusing Aligned and Non-Aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 48–57.
215. Goshvarpour, A.; Abbasi, A.; Goshvarpour, A. Fusion of heart rate variability and pulse rate variability for emotion recognition using lagged poincare plots. *Australas. Phys. Eng. Sci. Med.* **2017**, *40*, 617–629. [CrossRef]
216. Ghayoumi, M.; Thafar, M.; Bansal, A.K. Towards Formal Multimodal Analysis of Emotions for Affective Computing. In Proceedings of the DMS, Salerno, Italy, 25–26 November 2016; pp. 48–54.
217. Gao, Y.; Hendricks, L.A.; Kuchenbecker, K.J.; Darrell, T. Deep learning for tactile understanding from visual and haptic data. In Proceedings of the Robotics and Automation (ICRA), 2016 IEEE International Conference on IEEE, Stockholm, Sweden, 16–20 May 2016; pp. 536–543.
218. Dasdemir, Y.; Yildirim, E.; Yildirim, S. Emotion Analysis using Different Stimuli with EEG Signals in Emotional Space. *Nat. Eng. Sci.* **2017**, *2*, 1–10. [CrossRef]
219. Callejas-Cuervo, M.; Martínez-Tejada, L.; Botero-Fagua, J. Architecture of an emotion recognition and video games system to identify personality traits. In *VII Latin American Congress on Biomedical Engineering CLAIB 2016, Bucaramanga, Santander, Colombia, October 26th–28th, 2016*; Springer: Singapore, 2017; pp. 42–45.
220. Ringeval, F.; Eyben, F.; Kroupi, E.; Yuce, A.; Thiran, J.P.; Ebrahimi, T.; Lalanne, D.; Schuller, B. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognit. Lett.* **2015**, *66*, 22–30. [CrossRef]
221. Metallinou, A.; Wollmer, M.; Katsamanis, A.; Eyben, F.; Schuller, B.; Narayanan, S. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affect. Comput.* **2012**, *3*, 184–198. [CrossRef]
222. Haq, S.; Jackson, P.J.; Edge, J. Speaker-dependent audio-visual emotion recognition. In Proceedings of the AVSP, Norwich, UK, 10–13 September 2009; pp. 53–58.
223. Haq, S.; Jackson, P.J.; Edge, J. Audio-visual feature selection and reduction for emotion classification. In Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP'08), Moreton Island, Australia, 26–29 September 2008; pp. 185–190
224. Grimm, M.; Kroschel, K.; Mower, E.; Narayanan, S. Primitives-based evaluation and estimation of emotions in speech. *Speech Commun.* **2007**, *49*, 787–800. [CrossRef]
225. Pringle, H. *Brand Immortality: How Brands Can Live Long and Prosper*; Kogan Page Publishers: London, UK, 2008.
226. Kołakowska, A.; Landowska, A.; Szwoch, M.; Szwoch, W.; Wrobel, M.R. Emotion recognition and its applications. In *Human-Computer Systems Interaction: Backgrounds and Applications 3*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 51–62.
227. Li, G.; Wang, Y. Research on leamer's emotion recognition for intelligent education system. In Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chongqing, China, 12–14 October2018; pp. 754–758.
228. Majumdar, J.; Dhakal, P.; Rijal, N.S.; Aryal, A.M.; Mishra, N.K. Implementation of Hybrid Model of Particle Filter and Kalman Filter based Real-Time Tracking for handling Occlusion on Beagleboard-xM. *Int. J. Comput. Appl.* **2014**, *95*, 8887. [CrossRef]
229. Majumdar, J.; Aryal, A.M.; Rijal, N.S.; Dhakal, P.; Mishra, N.K. Implementation of Real Time Local Search Particle Filter Based Tracking Algorithms on BeagleBoard-xM. *Int. J. Comput. Sci. Issues (IJCSI)* **2014**, *11*, 28.
230. Smith, J.R.; Joshi, D.; Huet, B.; Hsu, W.; Cota, J. Harnessing ai for augmenting creativity: Application to movie trailer creation. In Proceedings of the 25th ACM international conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 1799–1808.
231. Mehta, D.; Siddiqui, M.F.H.; Javaid, A.Y. Recognition of emotion intensities using machine learning algorithms: A comparative study. *Sensors* **2019**, *19*, 1897. [CrossRef]
232. Jaiswal, S.; Virmani, S.; Sethi, V.; De, K.; Roy, P.P. An intelligent recommendation system using gaze and emotion detection. *Multimed. Tools Appl.* **2019**, *78*, 14231–14250. [CrossRef]