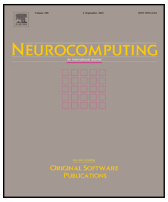
Neurocomputing 561 (2023) 126866

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Survey paper

A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods

Bei Pan, Kaoru Hirota, Zhiyang Jia [*], Yaping Dai

School of Automation, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Communicated by B. Wu

Keywords:
Emotion recognition
Multimodal information fusion
Feature learning
Classifier

ABSTRACT

Affective computing is one of the most important research fields in modern human–computer interaction (HCI). The goal of affective computing is to study and develop the theories, methods, and systems that can recognize, explain, process, and simulate human emotions. As a branch of affective computing, emotion recognition aims to enlighten the machine/computer automatically analyzing human emotions, which has received increasing attention from researchers in various fields. Human beings generally observe and understand the emotional states of one person by integrating the perceived information from his/her facial expressions, voice tone, speech content, behavior, or physiological features. To imitate the emotion observation manner of humans, researchers have been devoted to constructing multimodal emotion recognition models by fusing information from two or more modalities. In this paper, we provide a comprehensive review of multimodal emotion recognition from the perspectives of multimodal datasets, data preprocessing, unimodal feature extraction, and multimodal information fusion methods in recent decades. Furthermore, challenges and future research directions of the topic are specified and discussed. The main motivations of this review are to conclude the recent emergence of abundant works on multimodal emotion recognition and to provide potential guidance to researchers in the related field for understanding the pipeline and mainstream approaches to multimodal emotion recognition.

1. Introduction

Emotion is a significant component of human intelligence. Human behavior not only depends on rational thinking and logical reasoning but is heavily influenced by emotion. Since human society entered the information age, there has been considerable increase in human–computer interaction (HCI). In addition to adequate material, the primary needs of human beings also include satisfaction from the spiritual level. In order to achieve emotional HCI, computers are expected to possess the capacities of observing, understanding, and generating various emotions. Accordingly, these requirements give birth to the concept of affective computing [1].

Over the past two decades, researchers from psychology, neurophysiology, cognitive science, computer science, and other disciplines have been devoted to studying affective computing, which has already been applied to various fields. The applications can be roughly categorized into five classes: (1) Service industry. Robots in banking, hospitals, catering, government services, and other industries providing customers with affect services can improve customers' experience during or even the entire service process. (2) Education. By monitoring students' emotional states or concentration in class, robots can help improve the instructors' and students' teaching quality and learning efficiency, respectively. (3) Healthcare. Cognitive affective computing is integrated into the medical robot to assist the doctor with treating psychological diseases and provide emotional comfort for patients. (4) Intelligent driving. Emotion analysis and fatigue detection technologies are integrated into intelligent driving, which contributes to traffic accident prevention. (5) Entertainment industry. Adding emotion recognition and interactive technology to computer games can build more realistic virtual scenes, reduce player fatigue, and increase game entertainment.

Emotion recognition has received considerable attention as a significant component of affective computing. The aim is to discover the mapping relationship between external emotional representation and internal emotional state to identify the current type of human emotion. The distribution of emotion recognition publications is illustrated in Fig. 1, which is searched from Web of Science, Scopus, and Engineering Village. It is clear that an increasing number of emotion recognition-related methods have been reported in ten years. Human emotions can mainly be identified by facial expressions, speech, text, and physiological signals. Facial expression is the most straightforward way for human beings to convey emotions. Machines perceive facial expressions through processing and analyzing facial images or videos

[*] Corresponding author.
E-mail address: zhiyang.jia@bit.edu.cn (Z. Jia).

https://doi.org/10.1016/j.neucom.2023.126866
Received 31 January 2022; Received in revised form 15 April 2023; Accepted 30 September 2023
Available online 6 October 2023

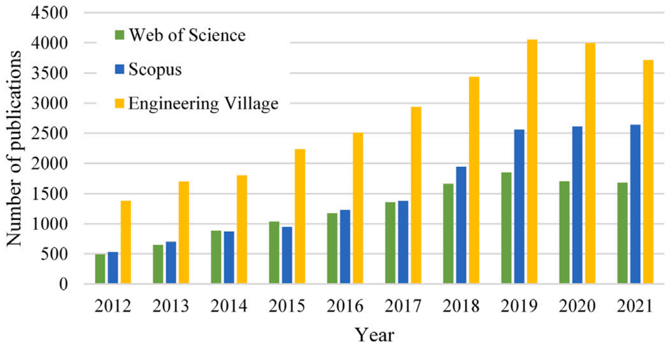


Fig. 1. Distribution of emotion recognition publications sorted by year. (Search string: (“affective” AND “computing”) OR (“emotion*” AND “recognition”) OR (“emotion*” AND “classification”) OR (“facial” AND “expression” AND “recognition”) OR (“sentiment” AND “analysis”)).

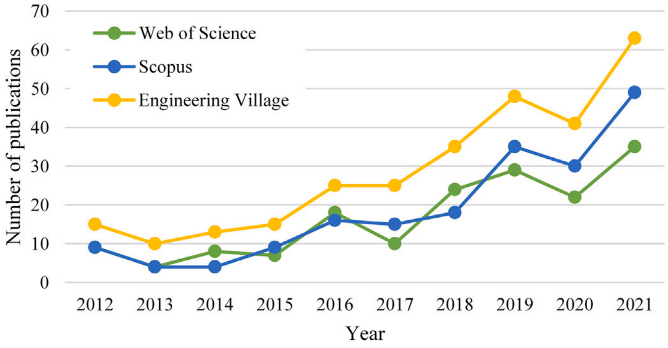


Fig. 2. Distribution of multimodal emotion recognition publications sorted by year. (Search string: (“multimodal” AND “affective” AND “computing”) OR (“multimodal” AND “emotion*” AND “recognition”) OR (“multimodal” AND “emotion” AND “classification”) OR (“audio*” AND “visual” AND “emotion”) OR (“bi-modal” AND “emotion” AND “recognition”) OR (“emotion*” AND “recognition” AND “fus*”) OR (“physiological” AND “emotion” AND “recognitio”)).

collected by cameras. Emotions delivered by speech signals contain not only the explicit and concrete content but also the vocal information of the speaker. Machines can capture emotional information from speech signals by analyzing the prosodic, voice quality features, and text information. Physiological signals, such as electroencephalogram (EEG), Electromyogram (EMG), and Electrocardiogram (ECG), are implicit emotional expressions. With the development of advanced non-invasive devices, emotion recognition based on physiological signals has received much research attention. Therefore, it is reasonable for machines to explore the relationship between physiological data and particular emotions.

Since emotions are expressed through multiple modalities, it is easy for human beings to perceive other person’s emotions or intentions by combining facial expressions, speech, or other information. To improve the performance of emotion recognition for machines, in the last two decades, much research in this field has been committed to fusing multimodal information for comprehensive and accurate emotion recognition. The distribution of multimodal emotion recognition publications searched from Web of Science, Scopus, and Engineering Village is illustrated in Fig. 2. It is apparent that there is an increasing interest in research on multimodal emotion recognition. The fusion of audio-visual, speech-text, audio-visual-text, visual-physiological, or multiple physiological modalities is a popular research direction. In this study, we review recent multimodal emotion recognition work from the perspectives of multimodal emotion datasets, data preprocessing, feature extraction of unimodality, and multimodal fusion emotion recognition.

We compare our paper with the published surveys on multimodal emotion recognition from five attributes in Table 1: (1) dataset, (2) data preprocessing, (3) feature extraction from different modalities, (4) fusion method, and (5) combinations of multiple modalities. It is clear from the comparison shown that our paper is distinguishable. Specifically, compared to most surveys on multimodal emotion recognition with wraparound or ignored introduction of data preprocessing, in this review, the data preprocessing methods of different single modalities are separately introduced. Furthermore, for the feature extraction methods, some of the recent popular deep learning-based technologies designed for emotional information extraction are represented. In addition, most existing reviews focus mainly on audio-visual multimodal fusion and introducing multimodal emotion recognition according to the fusion strategies. To make a clear presentation of the combinations of various modalities, we comprehensively introduce multimodal fusion from a new perspective of the fused modalities. Nine combinations of different modalities are contained and discussed, which contributes to learning the specific characteristics of individual modalities as well as the complementary of various modalities.

The rest of the paper is organized as follows. Emotion model and procedure are introduced in Section 2; Section 3 presents and lists several available multimodal emotion datasets; Data preprocessing methods are described in Section 4; Section 5 discusses feature extraction methods from unimodality; Section 6 provides the fusion methods of different modalities for emotion recognition. Finally, conclusions, challenges, and future work about multimodal emotion recognition are discussed in Section 7.

2. Emotion model and procedure

2.1. Categorization of emotion models

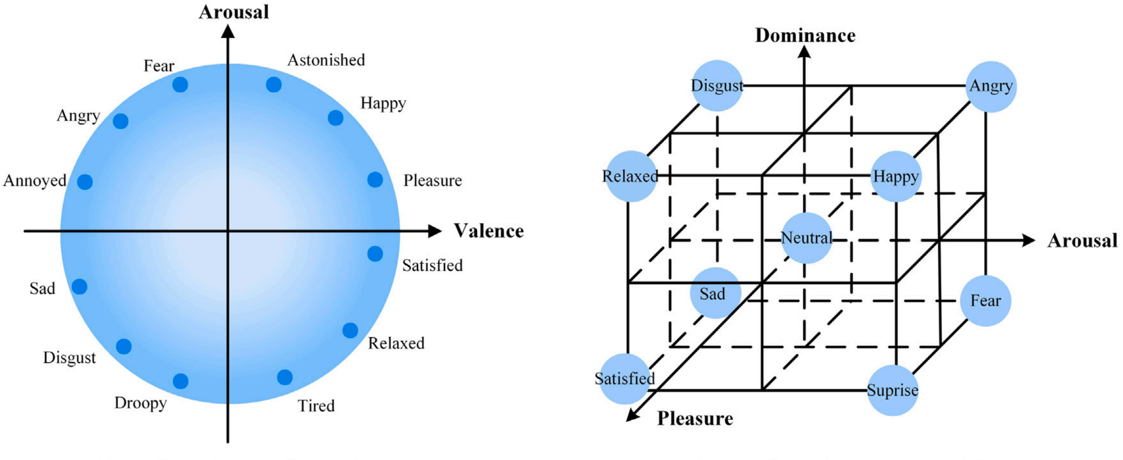
Generally, the type of emotion model can be categorized into discrete and dimensional representations. For the discrete emotion model, emotions are described as six basic categories, i.e., anger, disgust, fear, happiness, sadness, and surprise, which was defined by Ekman in 1971 [9]. These six basic emotions are universal across human ethnicity and cultures and can be used to compound other emotions. The primary emotions have the following characteristics: (i) emotions come from instinct; (ii) different people produce the same emotions under the same circumstances; (iii) different people express basic emotions similarly. One of the significant advantages of discrete emotion representation is that the categorical emotion scheme can describe people’s emotional experiences in daily life. Another is that it is intuitive to describe emotion based on the six emotion labels. Therefore, many efforts have been devoted to discrete emotion recognition.

An alternative emotion description is dimensional emotion. Some psychologists and experts in artificial intelligence consider that emotions can be represented through continuous dimensions. In contrast to discrete emotion, dimensional emotion theory defines different emotions as points in the dimensional space. The Valence-Arousal (VA) [10] and Pleasure-Arousal-Dominance (PAD) [11] are two typical and widely accepted dimensional emotion models. For the VA model, the valence dimension measures the positive and negative emotional states, while the arousal dimension indicates the intensity of emotions. PAD model adds the dominance dimension based on VA model, which defines as a feeling of control and influence over the surroundings and others. Two typical VA and PAD dimension emotion models are given in Fig. 3. It is noted that discrete and dimensional emotion representations can be transformed into each other to some extent.

Table 1
Comparison of the proposed review with the existing reviews in multimodal emotion recognition.

Review paper	Year	Dataset	Data preprocessing	Feature extraction of different modalities				Fusion method	Combinations of multiple modalities
				Audio	Video	Text	Physiological signal		
Zeng et al. [2]	2008	✓	✗	✓	✓	✗	✗	✓	A+V
Calvo and D'Mello [3]	2010	✗	✗	✓	✓	✓	✓	✓	A+V, MP
Wu et al. [4]	2014	✓	✗	✓	✓	✗	✗	✓	A+V
D'mello and Kory [5]	2015	✓	✗	✓	✓	✓	✓	✓	A+V, A+V+T, V+P, MP
Zhao et al. [6]	2019	✓	✗	✓	✓	✓	✓	✓	A+V, A+V+T
Jiang et al. [7]	2020	✓	✗	✓	✓	✓	✓	✓	A+V, V+T, A+V+T, V+P
Shoumy et al. [8]	2020	✓	✓	✓	✓	✓	✓	✓	A+V, A+V+T, V+P
This review		✓	✓	✓	✓	✓	✓	✓	A+V, V+T, A+T, V+P, V+B, MP, A+V+T, A+V+P, A+V+B

Legenda: A = Audio; V = Video; T = Text; P = Physiology; B = Body movement; MP = Multiple physiological signals.



(a) VA dimension emotion model

(b) PAD dimension emotion model

Fig. 3. Two typical dimension emotion models.

2.2. Procedures of emotion recognition

Both unimodal and multimodal signals can be involved in emotion recognition-related tasks to realize the emotion classification. Since we mainly focus on multimodal emotion recognition in this paper, the general procedures of emotion recognition based on multimodalities will be introduced. The pipeline of a typical multimodal emotion recognition system is presented in Fig. 4. [12]

i. data collection: induce subjects to generate emotions with the help of emotional stimuli, e.g., images, music, videos et al. collect and process all related emotional data to obtain emotion datasets.

ii. data preprocessing: remove noise or other interference for data in each modality. For example, detect and normalize the face in the visual modality.

iii. emotional feature extraction: design appropriate feature descriptors for learning representative features of different modalities.

iv. emotion recognition: design an effective fusion strategy to integrate features or classification scores of each modality for the final emotion decision.

3. Multimodal emotion datasets

The performance of the emotion recognition model depends highly on emotion data. Abundant labeled data are prerequisites for constructing an emotion recognition model with high performance and generalization. Emotional datasets are generally collected in a lab-controlled environment or in the wild. Emotions collected in the lab can be divided into two categories: acted and spontaneous emotions. The acted emotion datasets are recorded by asking participants to express different emotions, which are usually exaggerated behavior and beneficial for emotion classification. However, the acted emotions may hide the actual relationship between explicit expressions and implicit emotion states. Therefore, researchers have designed various environments and created emotion-induced materials to evoke and record the external and inherent responses. There are three main approaches to eliciting emotions: (1) Provide different videos, music, images, and other materials to elicit different emotions from subjects, which is the primary approach to collecting emotional data. (2) Design some simulated scenarios to recall the memories of the unforgettable emotional experience in the past life of the subjects. (3) Construct interaction scenarios for subjects, where they can talk about anything like some products, movies, songs, or their life experience to evoke emotions in each other. Compared with the acted ones, the evoked emotions are spontaneous, closer to the emotion expressed in the actual interaction, and conducive to practical application.

Compared with the lab-controlled environment, the emotion datasets collected in the wild are more natural and closer to real emotional states. Those datasets are generally collected from movies or websites. Specifically, the emotional states in video clips from movies or TV series are acted by professional actors in the wild environment, which can reflect real emotional states of individuals with different characters under a specific event or environment. It is noticed that there are various illumination variations, occlusions, face angle changes, and environment noises for videos in the wild, resulting in difficulties and challenges for high accuracy emotion recognition. Since numerous text comments for products, movies or events are uploaded on websites

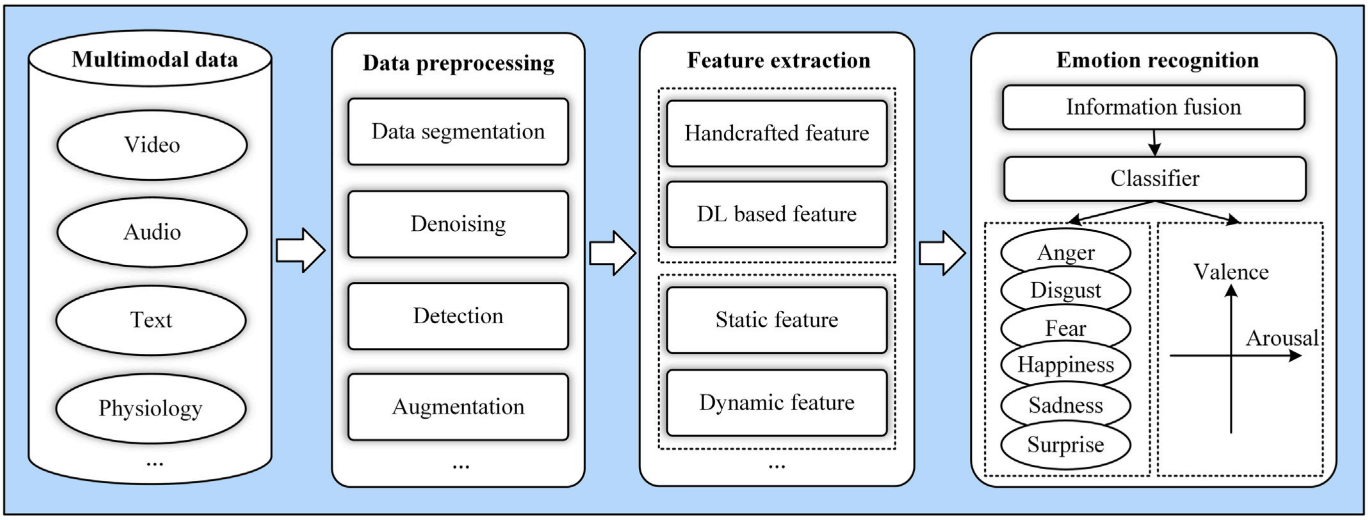


Fig. 4. The pipeline of a typical multimodal emotion recognition system.

Table 2
Multimodal emotion datasets.

Dataset	Subject	Size	Modality	Type	Emotion description	Access	Used in
eNTERFACE05 [13]	43	Available: 1290 audiovisual recordings	Audio and video	Acted emotions in labs	Six basic emotions, boredom, contempt, unsure, thinking, concentrating and bothered	http://www.enterface.net/enterface05/	[14–22]
RML [23]	8	720 audiovisual recordings	Audio and video	Acted emotions in labs	Six basic emotions	http://shachi.org/resources/4965	[18,24–28]
BAUM-1 [29]	31	Available: 1184 audiovisual recordings	Audio and video	Spontaneous emotions in labs	Six basic emotions	https://archive.ics.uci.edu/ml/datasets/BAUM-1	[24–26,30–32]
MELD [33]	304	13,000 utterances	Audio, video, and text	Spontaneous emotions in the wild	Six basic emotions, positive, negative and neutral	https://affective-meld.github.io/	[34–38]
IEMOCAP [39]	10	Approximately 12 h of data	Audio, video, and text	Spontaneous emotions in labs	Happiness, anger, sadness, frustration and neutral state categories; activation, valence and dominance dimensions	https://sail.usc.edu/iemocap/	[26,37,40–43]
SEMAINE [44]	150	959 conversations	Audio and video	Spontaneous emotions in labs	Five affective dimensions and 27 associated categories	https://semaine-db.eu/	[40,41,45]
RECOLA [46]	46	Available: 27 audiovisual recordings	Audio, video, and physiological signal	Spontaneous emotions in labs	Arousal-valence dimensions	http://diuf.unifr.ch/diva/recola	[47–49]
MAHNOB-HCI [50]	27	Available: 540 audiovisual recordings, 532 physiological and eye gaze data	Audio, video, physiological signals, and eye gaze	Spontaneous emotions in labs	Arousal, valence, dominance, predictability, and emotional keywords	https://mahnob-db.eu/	[51–53]
DEAP [54]	32	32 physiological signals and 22 videos	Video and physiological signal	Spontaneous emotions in labs	Arousal, valence, and dominance	http://www.eecs.qmul.ac.uk/mmv/datasets/deap/	[51,52,55–58]

such as YouTube and Facebook. It is available to collect those data for emotion analysis. The emotion datasets recorded in the wild are difficult to recognize but meaningful for real interaction and emotion-related application. In the following, several noteworthy multimodal emotion datasets listed in Table 2 are described in detail.

The eNTERFACE05 dataset. This is an audio-visual dataset constructed by Martin et al. [13] in 2006. The subjects were told to listen to six successive short stories, each eliciting a particular emotion. Then, subjects were asked to react to emotions of anger, disgust, fear, happiness, sadness, and surprise. Five reactions were simulated for each emotion. Two experts examined each recorded sample to decide whether or not each sample expressed the requested emotion unambiguously, and 42 subjects from 14 different countries were retained. The frame rate of each video sequence is equal to 25 frames per second, and the sampling rate of each audio sample is 48 kHz with the mono channel.

The RML datasets. This is an audio-visual dataset constructed by Wang et al. in 2008 [23]. Participants were asked to express their emotions as naturally as possible according to the provided emotional sentences, which were designed to recall the emotional incident that they had experienced. To enable the data to be used in a more general application, subjects speaking six languages, i.e., English, Mandarin, Urdu, Punjabi, Persian, and Italian, were invited to the data collection. To ensure that each subject expressed the expected emotions, at least two subjects who did not know the corresponding language were selected to test emotions. In addition, a video sample was added to the dataset while all subjects detected the intended emotion. Five hundred video samples were collected with six basic emotions. Each clip was recorded at a sampling rate of 22 050 Hz with mono channel 16-bit digitization.

The BAUM-1 dataset. This is an audio-visual dataset developed by Zhalehpour et al. [29] in 2016. 31 subjects spoken in Turkish were asked to watch a sequence of still images and short video clips, which were devised to evoke various emotions and mental states. They then had to express their feelings and ideas about the stimuli they watched on the screen without any guidance or scripts. The dataset contains six basic emotions, i.e., anger, disgust, fear, happiness, sadness, and surprise. In addition, several mental states are also contained, i.e., boredom, contempt, confusion, thinking, concentrating, bothered, and neutral. Five annotators were invited to annotate and give scores for each clip in the dataset. Finally, the majority voting over the five annotators gave each clip an emotion label.

The MELD dataset. This is a conversational dataset developed by Poria et al. [33] in 2018. It contains 13 000 utterances of 1433 dialogues from the TV series *Friends*. Each utterance involves data of audio, visual, and textual modalities. Three annotators were invited to look at the available video clip of the utterances, and the majority-voting approach was employed to give the final emotion and sentiment label of each utterance.

The IEMOCAP dataset. This is an audio-visual conversation dataset constructed by Busso et al. [39] in 2007. They designed two different approaches to induce and express emotions. One was based on a set of scripts that ten actors were asked to memorize and rehearse. Another approach asked the subjects to improvise based on hypothetical scenarios that were designed to elicit specific emotions. The corpus contains approximately twelve hours of data recording actors' face, head, and hand movements during scripted and spontaneous spoken communication scenarios. In the post-processing, the dialogs were manually segmented at the dialog turn level. Six annotators were invited to assess the emotional content of the dataset. Emotions of anger, disgust, excitation, fear, frustration, happiness, sadness, surprise, and neutral state were selected as annotations.

The SEMAINE dataset. This is an audio-visual dataset developed by McKeown et al. [44] in 2011. A Sensitive Artificial Listener (SAL) agent was built to engage a person in a sustained, emotionally colored conversation. The interactions contain two parties, a "user" and an "operator" (either machine or a person simulating a machine). Participants in the experiments interacted with two versions of SAL, one with the best nonverbal skills and one with a degraded set. There are 150 participants and 959 conversations recorded in the dataset. 6–8 annotators labeled per clip with five dimensions and 27 associated categories.

The RECOLA dataset. This is a multimodal corpus of spontaneous interactions in French constructed by Ringeval et al. [46] in 2013. Since this corpus was based on a study focusing on emotion perception during remote collaboration, 46 participants were equally separated into two teams. Each team participant was asked to solve the survival task individually in the separated rooms and received a questionnaire to evaluate their initial emotional state. A mood induction technique was used to balance the interaction context for the collaborative task. The data of audio, video, electrocardiogram (ECG), and electrodermal activity (EDA) modalities were recorded continuously and synchronously. In addition, six annotators measured the emotion on arousal and valence dimensions and social behavior labels on five dimensions.

MAHNOB-HCI. This is a multimodal dataset developed by Soleymani et al. [50] in 2012. Face videos, audio signals, eye gaze data, and peripheral physiological signals of 27 participants were recorded. Two experiments were conducted for data collection. In the first one, participants were asked to watch 20 inspirational videos and write down their emotion states using emotion labels and arousal, valence, dominance, and predictability. In the second one, they watched the short videos and images without tags first and then with correct or incorrect tags. Finally, participants assessed their agreement or disagreement with different tags.

The DEAP dataset. This is a physiological signals based emotional dataset developed by Koelstra et al. [54] in 2012. To elicit emotions, 40 one-minute-long excerpts of music videos were selected and played. The electroencephalogram (EEG) and peripheral physiological signals of 32 subjects were recorded while they were watching the inspirational videos. Besides, the frontal face video of 22 subjects was also recorded. Subjects valued each video regarding arousal, valence, like/dislike, dominance, and familiarity.

4. Data preprocessing

Data preprocessing is a fundamental step for multimodal emotion recognition. The primary purpose of data preprocessing is to eliminate irrelevant information, simplify data to the maximum extent possible, enhance the detectability of emotion-related features, and improve the reliability of feature extraction and recognition. For the multimodal dataset, data preprocessing is performed separately according to different modalities' data characteristics. Therefore, in the following sections, data preprocessing methods of the face image, speech signal, and physiological signal are separately introduced.

4.1. Face image preprocessing

Face detection, alignment, and normalization are three major tasks in face image preprocessing. Face detection is to locate the faces in the image and segment faces from the image according to the bounding box [59]. One of the classical and most used face detection methods is the boosted cascade of weak classifiers, proposed by Viola and Jones [60]. Deep learning-based methods are also developed for face detection, such as the cascaded convolution neural network (CNN) [61] and the discriminative complete feature-based CNN [62]. Face alignment is to rotate and frontal the detected faces to promise the in-plane consistency of different faces. The coordinates of facial landmarks are straightforward and effective for face alignment. Therefore, various landmark detectors, such as Active Appearance Models (AAM) [63] and multitask cascade CNN (MTCNN) [64], have been designed for accurate detection. Face normalization focuses on eliminating illumination variation, head pose, and other influences on facial expression recognition [65]. Histogram equalization, gamma intensity correction (GIC) [66], and homomorphic filter [67] are typical algorithms for image normalization. For pose normalization, it can be well solved by the generative adversarial network (GAN) [68].

4.2. Speech signal preprocessing

The original speech signals are non-stationary but can be seen as invariant in a short duration. Therefore, in the first phrase of preprocessing, the speech signal is framed into several segments with a length of 20 to 30 ms. Then, a window function is applied to each frame to reduce the energy leakage and obtain a signal closer to the natural spectrum. Hamming windows, rectangular windows, and Hanning windows are used in speech windowing. An utterance contains the voiced speech, unvoiced speech, and silence. The voiced speech reflects the speech activity during utterance and contains emotion-related information. It is necessary to detect the voiced signals and remove the unvoiced and silence frames [25]. In voice activity detection, zero-crossing rate, short-time energy, and auto-correlation are three general methods. The last step for speech signal preprocessing is noise elimination or reduction. For removal of background noise, spectral subtraction, minimum mean square error (MMSE), and log-spectral amplitude MMSE are commonly used methods [69].

4.3. Text preprocessing

One of the major tasks of text preprocessing is to separate a stream of characters into a set of word-like elements [70]. Characters like punctuation, white spaces, and emoticons should be careful attention during preprocessing. Specifically, it is common to remove punctuation characters before feature learning, which is crucial for the improvement of analysis speed and model performance. In emotion analysis, space is considered the boundary between words without having any meaning. Therefore, unnecessary spaces are usually removed in most cases. Emoticons are keyboard characters that depict some facial expressions, such as smile and frown. These characters are facilitate to distinguish emotion polarity and generally translated into corresponding words in an utterance. Other text preprocessing components such as conversion of capital letters, acronym expansion, spelling correction, and short-word removal are also important for the high performance emotion recognition.

4.4. Physiological signal preprocessing

The collected physiological data usually involve noise, interrupting the signal and hindering effective emotion-related feature extraction. Therefore, it is necessary to reduce noise for physiological signals before feature learning. High-frequency filter, low-frequency filter, notch filter, and Butterworth bandpass filter are commonly used for noise reduction [71]. Delta (0.5–4 Hz), theta (4–8 Hz), alpha(8–13 Hz), beta(13–30 Hz), and gamma(30–43 Hz) bands that contain emotional activity in the brain are filtered from different physiological signals, respectively, by adopting Butterworth filters. In addition, common methods such as principal component analysis (PCA), independent component analysis (ICA), and common spatial patterns (CSP) are employed to remove artifacts and noise for physiological signals.

5. Emotion feature extraction

Emotion recognition can be roughly categorized into two types according to the use of modality, i.e., unimodal and multimodal emotion recognition. Unimodal emotion recognition methods generally employ single channels, such as face images, speech signals, text, and physiological signals, to classify different emotional states. Multimodal emotion recognition uses two or more emotion channels to analyze emotion comprehensively. Emotion feature extraction is a significant part of both unimodal and multimodal emotion recognition because distinct features can facilitate precise results. The feature learning method for unimodal emotion recognition can be used for multimodal emotion recognition. Features are separately extracted according to the characteristics of single modalities in multimodal emotion recognition. Therefore, in the following subsections, feature extraction methods of different modalities are introduced and discussed in detail.

5.1. Facial expression features

The facial expression conveys essential information about the emotions, feelings, intentions, and physical states of others. Aiming to detect, analyze, and understand facial expressions of humans, automatic facial emotion recognition (FER) is a vital focus of basic research. Exhaustive surveys of facial expressions are published in [72–77]. In most multimodal emotion recognition studies, the facial expression is an essential component and plays a vital role in providing appearance information. One of the most critical tasks for FER and facial expression-related multimodal emotion recognition is to extract facial expression features for emotion classification. Feature extraction aims to learn and extract discriminative emotional information from images. The methods of facial expression feature extraction are separated as shallow learning-based and deep learning-based methods. Therefore, literature about various facial expression feature extraction will be briefly introduced according to the categories.

Shallow learning-based or handcrafted facial expression feature has been widely used for FER. Generally, handcrafted features can be separated into three categories: geometry, appearance, and motion [78]. Geometric features refer to the track of facial landmarks, which can be presented as fiducial points, face mesh, active shape model changes in displacement between points around the eyes and mouth [79–82]. Appearance feature represents changes in skin texture and representative appearance feature descriptors are local binary pattern (LBP) [83], histogram of gradients (HOG) [84], Gabor wavelets [85,86], et al. Motion feature includes optical flow [87] and motion history images (MHI) [88].

The above-mentioned feature descriptors are designed to extract static features from a single image, while the dynamic information along the temporal domain is lost. Several dynamic feature extractors were developed to better understand the perception of dynamic changes in facial expression to capture spatial–temporal features from image sequences with ordered frames. For instance, the volume local binary patterns (VLBP) that combine the motion and appearance, the local binary pattern on three orthogonal planes (LBP-TOP) [89], and the histogram of oriented gradient from three orthogonal planes (HOG-TOP) [90]. As an extension of the static texture descriptor, dynamic events on the face are well tracked and calculated through these spatial–temporal texture descriptors. In [91], dynamic information from geometric and texture features is exploited through a landmark tracking framework and a parametric space. Their experiments have verified that, compared with the static features, dynamic features improve facial expression recognition and are more robust to some uncontrolled variations in video sequences.

Considering that in real scenarios, various disturbances are attached to the collected video sequences, such as pose variation, subject difference, and dynamic background. Efforts have been made to learn robust and generalized facial expression representations. Sariyanidi et al. [92] designed a dynamic feature extraction framework that represents facial expression variations as a linear combination of localized basis functions. The designed facial expression representation can not only recognize expressions with different intensities but deal with the temporal inconsistencies that existed in varying datasets. In [93], authors proposed a dynamic kernel-based representation that assimilates facial movements captured using local spatio-temporal representations in a large universal Gaussian mixture model (uGMM). With dynamic kernels, local representations from various parts of the face are acquired, and dynamic changes of different expressions are distinct for classification.

With the advancement of deep learning (DL) [94,95], it has been widely used for deep facial feature extraction because of the strong feature learning ability [96–100]. By using hierarchical architectures, DL-based facial expression feature descriptors can extract high-level emotional information from images [101]. In facial expression feature learning, CNN is often used to extract local facial features by using a set of filters to identify important patterns or features. Generally, the performance of DL-based methods relies heavily on the scale of the dataset. To obtain high-level discriminative features, abundant and diverse facial expression training data is required. However, most of the datasets used for FER have small scales, which is hard for efficient DL model training. To solve this issue, transfer learning is applied to FER by fine-tuning parameters in the pre-trained deep CNN model, such as AlexNet, ResNet, and VGG, which are trained on the large-scale datasets [102,103]. Fine-tuning some of the layers of the pre-trained model can not only decrease the computation cost but obtain distinct facial expression features.

In recent years, GAN has been widely used in facial expression recognition tasks for data augmentation and reducing the negative influences of emotion-unrelated variables. In [104], the GAN-based framework was designed to expand the training dataset and disentangle the expression, identity, and pose from an image to facilitate discriminative feature extraction. Similarly, in [105], a GAN-based structure guided by geometry information was proposed to produce identity-preserving face images with different poses and expressions. The training dataset is enlarged by generating labeled facial expression images, enabling the model to learn features of different emotions efficiently.

With the capability of finding salient regions in the image, attention mechanisms have been embedded in DL models to weight features according to their significance in facial expression recognition task [106]. To reduce the influences of head pose variation, occlusion, and low image resolution, Liu et al. [107] used the "visual attention" mechanisms to learn deep distinctive expression features based on saliency-guided facial patches from images in an unconstrained environment. Attention-based methods focus on learning salient features from images collected in the wild, which can suppress the influences of unrelated variables and increase the discrimination of facial expression features.

To model the dynamical evolution of facial expressions, recurrent neural network (RNN), long short-term memory (LSTM), gated recurrent unit (GRU) and 3-dimensional CNN (3D-CNN) have been applied to learn temporal relationships for the ordered image sequences [108,109]. Furthermore, numerous spatial–temporal feature extraction frameworks have been constructed to learn spatial features from single images and capture dynamic changes of frames [110,111]. Considering that not all frames are equally crucial for spatial–temporal feature extraction, in [112], a frame attention network was proposed to adaptively assemble the frame features to form a single distinct representation. For single-frame features, self-attention weight is firstly learned by a fully connected layer and softmax function. Then, it is refined by modeling the relation between two frames.

It is a challenging problem that most facial expression features are extracted and evaluated on the same distribution database, which may make it difficult to classify the features distributed on different domains. To solve the cross-domain problem, Zong et al. [113] proposed a transductive transfer regression model (TTRM) to bridge the feature distribution gap between the source and target domains. TTRM can obtain the discriminative expression features with the capability of quantifying contributions of different facial local regions. By analyzing the influence of learning complexity, Xia et al. [114] proposed a recurrent convolutional network (RCN) to explore the shallower architecture and lower-resolution input data. Chen et al. [115] constructed an adversarial graph representation adaptation (AGRA) framework to accomplish effective cross-domain local–global feature co-adaptation. As a result, domain-invariant features and more detailed content are obtained for distinguishing different expressions.

5.2. Speech emotion features

Speech is a natural and effective medium to express emotional states. Various studies have been conducted to extract emotional information and predict and analyze human emotion from speech signals [116–118]. It is worth noticing that speech signal, often coupled with facial expression or text, is a vital component in multimodal emotion recognition. Acoustic feature extraction is a significant and challenging task in speech emotion recognition (SER) and speech-related multimodal emotion recognition. Prosodic, voice quality, and spectral features, called handcrafted acoustic features, have been intensively explored and used for SER. In addition, with the advantages of strong learning ability, DL has been applied for exploring discriminative deep speech emotion features [119,120]. For example, CNNs are often employed to extract local acoustic features, such as variations in frequency content over time, from speech signals by analyzing the spectrogram of the signal.

In general, emotional speech data expressed by different speakers demonstrated large variations in acoustic characteristics, even if they intend to express the same emotion. Therefore, it is significant to extract discriminative and speaker-independent features that do not rely on speakers [121–124]. In [125], a two-layer fuzzy multiple random forests was proposed to extract speech emotion features. Non-personalized features are obtained through the derivation of basic acoustics features and fused with personalized features. In [126], a subject-independent method was developed using voice features from OpenSmile toolbox and higher-order spectral features. A particle swarm optimization-assisted biogeography-based optimization (PSOBBO) algorithm was designed to remove irrelevant and redundant features. Such a feature selection strategy can facilitate acquiring salient speech emotion features.

Similar to image frames in a video clip, dynamic information among speech segments in an utterance is vital for distinguishing emotions. Therefore, to better classify speech emotions, it is significant to model the temporal dependency among segments and obtain dynamic speech representations. With the capability of capturing context information, LSTM, often combined with CNN, is explored in various studies. Generally, CNN is used to extract features from segments, and the LSTM network is employed to capture temporal dynamic dependency of the segment features [127,128]. According to the experimental results, these studies showed that the cooperation of CNN and LSTM can increase the discrimination of speech features compared with using CNN alone. Recently, attention mechanism-based networks have been developed for speech feature learning [129,130]. In [131], a self-attention mechanism was added to bi-directional LSTM (BLSTM) to calculate the similarity between two frames and automatically assign weights to frames. The introduction of the attention mechanism facilitates the network in finding the salient emotion components and learning representative features.

In most cases, SER achieves well performance when the distribution of the testing set is close to the training set. However, unsatisfied recognition results may be obtained if training and testing sets have different distributions. To solve this issue, several researchers have been devoted to investigating the cross-corpus, and cross-language learning capable of taking multiple datasets at once and constructing a more robust SER model [132–135]. Domain adaptive representation learning [136] and domain adversarial neural networks [137,138] are popular in cross-corpus and cross-language SER. The former attempts to minimize the difference between the source and target domains. The latter focuses on learning common feature representations. Song and Zheng [139] proposed a framework called feature selection-based transfer subspace learning (FSTSL) to learn a robust low-dimensional corpus-invariant feature representation and improve the generalization of the cross-corpus SER model. Further, Song [140] developed a transfer linear subspace learning (TLSL) framework to learn a common feature subspace for source and target corpora. In these studies, robust corpus-invariant feature representations are obtained that improve the discrimination of cross-corpus speech emotions.

5.3. Text features

Textual information coupled with acoustic features in speech plays a vital role in multimodal emotion recognition. Grammatical analysis and semantic analysis are two important contents for textual feature extraction. Similar to facial expression and speech modalities, textual feature representation methods can be roughly categorized into traditional and DL-based methods. Bag of Words (BOW), rule-based technique, and statistical methods are typical traditional textual feature extraction methods [141]. Based on the emotion lexicons, Jin et al. [142] designed a textual emotion representation called eVector, which is combined with BOW to obtain lexical feature representations.

Due to the strong capability for feature learning, DL has been widely employed to learn emotional representations of text. Su et al. [143] proposed to extract emotional word vectors from the word2vec model and adopted an autoencoder to acquire the bottleneck features. The final textual features were obtained by concatenating the features in the semantic word vector and the bottleneck features together. Liu et al. [144] proposed a CBOW method that combines BOW and CNN for learning textual emotion features. Specifically, a CBOW model based on a feedforward neural network was first designed for vector representation of text, and then a CNN was trained for semantic features learning. LSTM is widely used for distinct emotion feature extraction from textual information to learn the contextual dependency in an utterance. Wang et al. [145] proposed a tree-structured CNN-LSTM model that extracted both local and global information from sentences. They trained the CNN model to learn the conducive emotion information from the divided text regions instead of a whole text. Then, the regional information was sequentially integrated by LSTM to learn long-distance dependencies among the whole sentence. Huang et al. [146] designed an emotion-enhanced LSTM (ELSTM) to improve the feature learning ability of LSTM by introducing emotional intelligence and attention mechanism.

5.4. Physiological signal emotion features

The above-mentioned explicit emotion signals are subjective in emotion expression. Compared with facial expression, speech, and text, physiological signals are implicit and objective, which are difficult to conceal deliberately. Therefore, much attention has been paid to physiological signals-based emotion recognition [12]. The common physiological signals [147] in emotion recognition are listed in Table 3. Among these signals, EEG is the most frequently used in physiological signals-related emotion recognition [71]. Conventional EEG features are categorized into time-domain, frequency-domain, and time-frequency domain [148,149]. Hjorth feature, fractal dimension feature, and higher order crossing feature are representative time-domain features. Frequency-domain features are usually extracted from five frequency bands, i.e., delta, theta, alpha, beta, and gamma, like power spectral density (PSD), differential entropy (DE), and higher order spectra (HOS). Time-frequency features exploit both temporal information and frequency domain information and help to explore the dynamical changes. The frequently-used time-frequency domain approaches contain wavelet transform, Hilbert Huang transform (HHT), and wavelet packet transform.

It is noted that the conventional features extracted from different domains have been independently used or integrated for EEG emotion analysis according to the specific task or data. In recent studies, DL-based methods have been gradually introduced into EEG feature extraction. For instance, CNN is commonly used to extract relevant spatial and temporal features from the EEG signals. Generally, raw EEG signals or conventional features are fed into various deep neural networks, such as deep belief network (DBN), CNN, RNN, and graph neural network (GNN) for high-level EEG representations learning [150,151]. As EEG signals are collected from different electrodes that record the energy changes over the scalp, the position of each electrode provides extra spatial information. To achieve a comprehensive analysis of brain activities during emotion generation, in most cases, electrode positions are projected into 2-D plain and treated as spatial information for spatial–temporal feature extraction [110,152]. Considering the biological topology of the human brain, there are internal connections among EEG channels and it is necessary to exploit the inter-channel relations to reveal the emotion-related functional connectivity. With the advantages of topological structure, GNN and its variants have been developed to capture the intrinsic relationship among EEG channels for discriminative emotion feature extraction [153,154].

EEG signals collected from various electrodes on the scalp comprise a variety of brain activity information, while not all information is necessary or equally important for EEG emotion recognition. Hence, it is important to investigate the correlation of different channels or frequency bands to emotional brain activities. Approaches such as group sparse canonical correlation analysis [155], time-frequency and weight distributions analysis [156], and attention mechanism [157] have been used for importance analysis of channels and frequency bands. Researchers have found that the frontal, parietal, and occipital regions are mostly related to emotion processing in the brain. Besides, higher frequency bands i.e., delta and gamma, are discriminative for emotion recognition. In [158], visualized activation maps were shown to prove these findings. Except for the discoveries of emotional-related brain regions and frequency bands, they also identified the global inter-channel relations between the left and right hemispheres can provide helpful information for emotion recognition. Taking these discoveries into consideration during feature learning helps to decrease computation costs and increase feature extraction efficiency. EEG signals are sensitive to subjects, meaning that a large difference in EEG exists among individuals. To solve the performance degradation under the cross-domain situation, multisource transfer learning [159], domain-invariant feature extraction [157], and transferable attention neural network [160] have been developed and are facilitate recognize emotions from new subjects.

5.5. Discussion

In this section, the popular methods of feature extraction for different modalities are discussed. Although the data types of different modalities are various, it can be concluded that there is similar feature extraction mechanism for these modalities. Firstly, handcrafted (or shallow) and DL-based (or deep) features are common categories in these modalities. Researchers designed varieties of handcrafted feature descriptors for each modality according to the characteristics of data and emotion expression. The handcrafted feature extraction methods possess the advantages of easy understanding, simple computation, and low memory consumption. Although more and more powerful DL-based methods have been proposed and applied in deep emotion feature extraction. For some modalities, such as speech and EEG signals, handcrafted features are still fundamental and play a vital role in DL-based high-level semantic feature learning.

Recently, deep neural networks, such as CNN and RNN, have been successfully utilized for discriminative emotion feature extraction in all modalities. For high-level representative feature learning, speech and EEG signals are displayed as images and fed into a deep model to take advantage of CNN’s hierarchical learning. Except for learning spatial features using CNN, in different modalities, 3D-CNN, CNN-LSTM, and graph convolutional network are developed for spatial–temporal features learning from sequential data. Due to its ability to focus on key parts, attention mechanisms are used to identify emotional regions and channels during the feature extraction process. It is also possible to transfer certain attention modules from one modality to another. Additionally, the feature variation cross-domain has been taken into account when extracting features from different modalities. There have been a number of domain adaptation strategies developed to handle the variations between the source domain and the target domain.

Table 3
Physiological signals for emotion recognition.

Name	Description
Electroencephalogram (EEG)	A direct reflection of brain activity and contains meaningful psychophysiological information for emotion recognition.
Electromyogram (EMG)	It reflects the functional status of facial muscles, where Corrugator muscles and Zygomaticus muscles are separately sensitive to positive and negative emotions.
Electrooculogram (EOG)	It can be used to measure the vertical and horizontal movement of eyes that provide useful information for valence recognition.
Electrocardiogram (ECG)	A measurement of beat-to-beat temporal changes of the heart rate, which gives deep insight to the emotional system of the human body.
Heart Rate Variability (HRV)	The most natural choice for arousal detection using comparison of sympathetic and parasympathetic frequency bands of the time series.
Galvanic Skin Response (GSR)	Measure the skin conductivity that decreases during relax states and increases when exposed to effort.
Skin Temperature (ST)	It can be used to identify if a person is relaxed or not.
Respiratory Rate (RES)	It can reflect multiple emotion states. For example, deep and fast breathing can indicate happiness or anger, irregular respiration patterns are a sign of negative valence and arousal.
Blood Volume Pressure (BVP)	It is described by the pulse-wave of the heart and the volume of the blood flowing through a vessel.

6. Multimodal emotion recognition

Multimodal emotion recognition has attracted increasing attention in affective computing [161]. To better understand human emotions for computers, it is necessary to imitate the way humans observe emotions. Humans judge other person's emotions by synthetically analyzing the information presented from facial expressions, voice, the content of utterances, and gestures during an interaction. Therefore, collecting emotion expression-related information and fusing different modalities help computers comprehensively recognize emotions. It is noticed that although the data forms of different modalities are heterogeneous, the inner consistency of semantic enables the collaboration of multiple modalities to acquire a more convinced emotion recognition model [162]. Specially, in [163], authors discussed when and why multimodal outperforms unimodal jointly through a theoretical treatment. In the previous sections, we have discussed the emotional feature extraction methods of individual modalities. In this section, state-of-the-art methods for multimodal emotion recognition will be discussed in detail.

Video, audio, text, and physiological signals are frequently used modalities in multimodal emotion recognition. In addition, various combinations of individual modalities have been put forward in the last decade, such as audio and video, speech and text, and multiple physiological signals fusion for emotion recognition. The fusion methods are generally categorized into feature-level, decision-level, and hybrid fusions. Feature-level fusion firstly concatenates features from different modalities and then trains a classifier for emotion classification. In decision-level fusion, classifiers of different modalities are trained separately, and the classification results are fused for the final decision. The hybrid fusion integrates the feature- and decision-level fusion. An illustration of multimodal fusion methods for emotion recognition is shown in Fig. 5, where the audio, video, and text modalities are chosen as an example.

6.1. Classifier

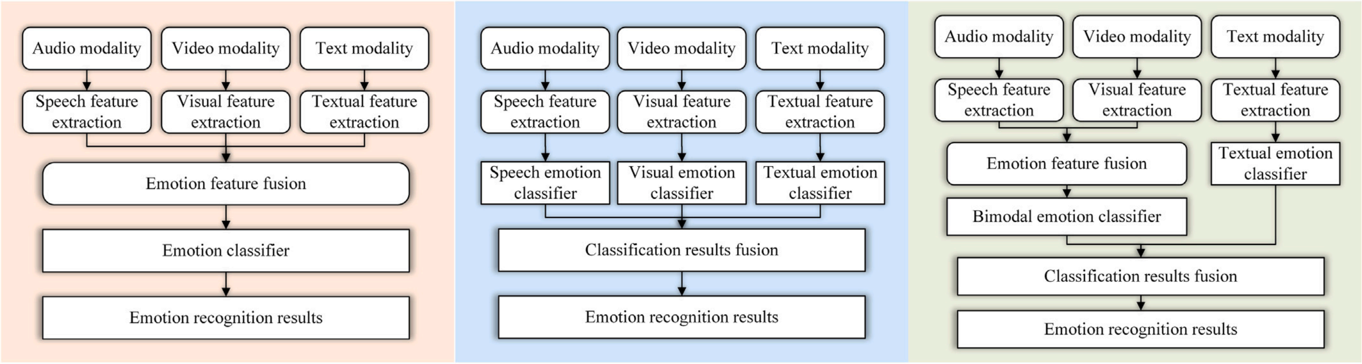
The classifier that decides the underlying emotion plays an essential role in multimodal emotion recognition. Various classifiers have been implemented for emotion recognition, such as support vector machine (SVM), random forest (RF), artificial neural networks (ANN), and many others. Due to the fact that each classifier has its own advantages and limitations, it is difficult to decide which is the most appropriate for emotion classification. Therefore, researchers usually choose or design emotion classifiers according to the specific task or characteristics of emotion features. Generally, a classifier's feature separation ability dramatically influences emotion recognition performance. As the most used classifier in emotion recognition, SVM aims to locate hyperplanes that accurately separate various feature groups. As most emotional features cannot be separated linearly, SVM involves kernel functions, such as linear, polynomial, and Gaussian, to convert features into a high-dimensional space for linear separation. In DL-based emotion recognition, high-level features are separated through a loss layer at the end of the network. The loss function greatly influences the classification accuracy. The softmax loss that minimizes the cross-entropy of the predicted probability and the truth ground distribution is the most used function in CNN models [75,164].

6.2. Audio-visual emotion recognition

Audio and visual modalities are the two most commonly used modalities in the multimodal emotion recognition research community. Audio modality possesses important information relevant to the intensity of emotions. In comparison, visual modality depicts the expressions of an image sequence, which contains rich appearance information. These two modalities contain almost 90% explicit emotional information and complement each other. Therefore, it is beneficial for performance improvement of emotion recognition through the integration of these two modalities. Numerous studies have focused on fusing emotional information of audio and visual modalities. This section introduces the representative audio-visual emotion recognition methods in recent years according to three fusion strategies. Comparisons of different methods fusing audio and visual information for emotion recognition are list in Table 4.

6.2.1. Feature-level fusion

Since a deep belief network (DBN) is able to learn the high-level relationship of input data effectively, it has been used to fuse the audio and visual features and learn feature representation. In [25], voice activity detection (VAD) was executed to distinguish whether an audio frame is salient and assigned weights of 0 or 1 for the salient and voice frames of the corresponding facial expression frames. Emotional features of facial and voice were extracted from the pre-processed data by CNN models and then fused by the DBN. While the average accuracy of single modalities was 69.8% on eNTERFACE dataset, the proposed fusion method produced a significantly higher accuracy of 85.69%. Zhang et al. [24] proposed a deep hybrid model to bridge the emotional gap between emotions and audio-visual features. The audio and video signals were first segmented into fixed frames, and then the CNN and 3D-CNN were employed to learn audio and visual segment features, respectively. Finally, the output features were concatenated and put into a DBN for learning a discriminative audio-visual segment feature representation. Pini et al. [15] proposed a multimodal DL architecture composed of three sub-networks. The 2-dimensional (2D) and 3D networks were developed for static facial features and


Audio modality
Video modality
Text modality
Speech feature extraction
Visual feature extraction
Textual feature extraction
Emotion feature fusion
Emotion classifier
Emotion recognition results
Audio modality
Video modality
Text modality
Speech feature extraction
Visual feature extraction
Textual feature extraction
Speech emotion classifier
Visual emotion classifier
Textual emotion classifier
Classification results fusion
Emotion recognition results
Audio modality
Video modality
Text modality
Speech feature extraction
Visual feature extraction
Textual feature extraction
Emotion feature fusion
Textual emotion classifier
Bimodal emotion classifier
Classification results fusion
Emotion recognition results

(a) Feature-level fusion

(b) Decision-level fusion

(c) Hybrid fusion

Fig. 5. Multimodal fusion methods for emotion recognition.

dynamic patterns. LSTM was applied for the audio branch to capture the temporal evolution of the audio features. The features trained from three networks were concatenated and sent into a fully connected layer for the final classification.

Similarly, in [17], audio and visual features were separately learned through the 2D and 3D CNN models. A non-linearity fusion approach with an extreme learning machine (ELM) was employed for fusing features obtained from two CNN models. They designed the two-stage ELM containing two ELM models. The first ELM was trained for gender separation, and the last layer was removed after training. The hidden layer of the first ELM was fed into the second ELM for emotion classification. In such a manner, emotions were recognized based on gender, which inherently eliminates the influence of gender and increases accuracy. To exploit the dependencies and relations of different modalities, Ghaleb et al. [14] put forward a metric learning method to jointly obtain a discriminative score and a robust representation in a latent space. The developed framework was scalable because it learns modality-specific metrics without imposing any constraints. Additionally, the rationale of the proposed distance was intuitive, which is beneficial for model explanation according to the importance of a single modality. Kansizoglou et al. [26] proposed an online autonomous emotion recognition paradigm to exploit both facial and voice emotional information. To distinguish different emotion states under time variations, the fused features were proceeded to an LSTM layer, which was monitored by a reinforcement learning agent. Audio and visual features were fused by a deep neural network (DNN). The unimodal features produced an average accuracy of around 44%, while the fused features achieved around 58% on the BAUM dataset.

Nguyen et al. [20] developed a generic framework that cascades 3-dimensional CNN (C3Ds) and DBN to extract spatio-temporal features from audio and visual modalities. They adopted the multimodal compact bilinear pooling (MCB) to capture the complex and intrinsic associations between the two modalities. The unimodal feature set produced an average accuracy of 83.09%, while the fusion of audio and visual achieved an increase of around 7%. Additionally, accuracy increased from 89.39% to 90.85% when MCB was applied, which further supported the efficacy of the fusion strategy. Huang et al. [165] used the transformer model to fuse the audio and visual modalities. In the framework, multi-head attention was employed to produce intermediate representations of multimodalities from a common semantic feature space. They found that the order of audio and visual features impacts recognition accuracy. Results showed that audio on the left and video on the right improve performance because visual features are captured as principle parts. This observation further indicates that visual feature is superior to audio feature in emotion recognition.

These methods focus on fusing audio-visual features with neural network-based approaches. High-level discriminative fused representations are generated, but the specific characteristics of single modalities are lost. In [47], a cross-attentional fusion approach was introduced to encode the inter-modal information and preserve the intra-modal characteristics. This fusion strategy explored the inter-modal relationships by computing cross-attention weights of audio-visual modalities. The salient feature representations from two modalities were combined and fed to the fully connected layers for the valence and arousal predictions. The average accuracies of valence and arousal of single modalities were 55.25% and 70.2% on the RECOLA dataset, while the proposed model achieved significantly higher accuracies of 69% and 83.8%, respectively.

6.2.2. Decision-level fusion

In [167], genetic algorithms (GA) was implemented to learn the most suitable HMM structure for the speech system. For the vision system, PCA was employed for dimensionality reduction and emotion identification. GA was also designed to optimize ANN's structure, improving the vision system's performance. A weighted sum of recognition probabilities of speech and vision systems produced the final recognition results. Noroozi et al. [18] proposed to define a set of key-frames of each video by k-means cluster. Visual features were described using geometry deformation and a CNN-based model from the selected key-frames. An 88-dimensional feature vector was computed with the set of audio features of pitch, intensity, and Mel-frequency cepstral coefficient (MFCC) et al. SVM classifiers were trained for each modality, and the obtained confidence outputs were used to define a new feature space for the final prediction. Fan et al. [168] developed a hybrid network combining CNN-RNN and C3D to simultaneously model appearance and motion features. SVM model was trained for audio modality, and the prediction results of the three models were combined through a weighted summation.

In [169], three networks were separately designed to learn the emotional features from visual and audio modalities. The spatio-temporal texture features and dynamic geometric features were separately extracted from the C3D and CNN-LSTM networks. The acoustic features that could complement the limitations of image-based networks were extracted from an audio-based network. Further, they proposed an emotion adaptive fusion strategy by measuring the recognition accuracy per emotion through a given validation dataset. As a powerful tool to improve classification performance, ensemble learning has been applied in decision-level fusion emotion recognition [170]. Conventional ensemble rules include max, min, sum, average, and product, which are used for combining several classifiers. In [171], with simple weighted averages, authors presented a new way of aggregating models based on random hyperparameter searches. In [16], both manual and deep features were learned separately from audio and visual modalities. They designed the blending ensemble algorithm to fuse the classification results from SVM and multi-task CNN classifiers for the final emotion decision. This fusion method has achieved a significantly higher accuracy of 81.36% than both audio (56.33%) and visual (66.93%) modalities. From these methods, it is concluded that

Table 4
Recent work on audio-visual fusion emotion recognition.

Reference	Preprocessing	Feature extraction	Fusion strategy	Classifier	Dataset	Experimental condition	Accuracy
[21]	Audio signal processing; face detection and alignment	Prosodic, MFCC, formants features of audio modality, QIM and ITMI of visual modality	Hybrid-level: weighted averaging and stacked generalization	MLP RBF	eNTERFACE05	LOSO	77.78%
[24]	Overlapped data segment	CNN for segment audio features extraction, 3D CNN for segment visual features extraction	Feature-level: DBN	SVM	RML BAUM-1s eNTERFACE05	CV LOSO CV LOSO CV LOSGO	80.36% 54.57% 85.97%
[15]	Audio signal sampling, face detection and alignment, data augmentation	Temporal audio features extracted by LSTM, static and dynamic visual features extracted by 2D and 3D CNN	Feature-level: concatenation	ANN	AFEW	V_{train} = 774, V_{val} = 383, V_{test} = 653	49.92%
[18]	Key frame selection	88-dimensional audio features, geometric and CNN-based visual features	Decision-level: stacking fusion	SVM RF	SAVEE eNTERFACE05 RML	CV 10F	100% 99.72% 98.73%
[166]	Face detection, facial landmark annotation, data augmentation	CNN for audio features extraction, CNN-BRNN for facial texture extraction, SVM and CNN for facial landmark action features extraction	Hybrid-level: concatenation, weighted summation	SVM	AFEW6.0	CV 3F	56.66%
[20]	Face detection	3D CNN for spatio-temporal audio and visual feature extractions	Feature-level: multimodal compact bilinear pooling	DBN	eNTERFACE FABO	CV LOSO CV LOSO	90.85% 92.24%
[14]	Face detection and alignment	openSMILE for audio features extraction, CNN for visual features extraction	Feature-level: metric learning	SVM	eNTERFACE05 CREMA-D	CV 10F CV 10F	66.5% 91.5%
[26]	Speaking time detection, face detection, data augmentation	VGGish models for audio and visual features extraction	Feature-level: DNN	ANN	RML BAUM-1s	CV LOSO CV LOSGO	82.97% 56.01%

(*continued on next page*)

compared with feature-level fusion, the decision-level fusion strategy preserves the characteristics of single modalities but ignores the inter-modality relationship.

6.2.3. Hybrid fusion

Hybrid fusion combines the advantages of feature- and decision-level. The unique attributes of different modalities are retained, and the correlations among modalities are explored. However, the limitation is that the model complexity may increase. Bejani et al. [21] simulated human emotion states by combining facial expression and speech emotional information. Prosody feature, MFCC, and formant frequencies were extracted as vocal features. Integrated time motion image (ITMI) and quantized image matrix (QIM) images were used for facial expression feature extraction. Then, audio- and visual-based emotion classifiers were trained separately for unimodal emotion recognition. To realize feature fusion, all features extracted from two modalities were cascaded and then selected through analysis of variations. Final recognition results were obtained by combining the benefit of feature- and decision-level fusions. This hybrid fusion method produced an accuracy of 77.78% on the eNTERFACE dataset, which is a significant improvement compared with using audio (54.99%) or visual (39.27%) modalities alone. In [166], three complementary cues, i.e., facial landmark action, facial texture, and audio signal, were explored and fused for emotion recognition from video clips collected in the wild. The hybrid fusion was applied in the multi-cue fusion framework, where dynamic facial features were concatenated for feature-level fusion, and classification results from visual and audio were integrated for decision-level fusion. Their results showed that emotion recognition performance in the wild is improved by integrating multiple cues from audio-visual modalities.

6.3. Audio, visual and text modalities fusion

In [172], Deep Boltzmann Machine (DBM) was applied to learn a joint density model over the input of visual, audio, and textual modalities. The deep architecture of DBM enlightened the possibility of discovering the highly non-linearity between low-level features and the complex relationship of different modalities. In [173], Poria et al. developed a multimodal information extraction agent, which adopted an ensemble feature extraction approach by exploiting the joint use of audio, visual, and text modalities information. A feature-level fusion strategy was utilized by concatenating the tri-modal features together. In their further work [19], CNN was integrated with a low-dimensional RNN to capture spatial structure information in static images and temporal patterns inherent in a video sequence. Besides, multiple kernel learning was developed to adaptively combine emotion features in audio, video, and text. For single modalities, the best accuracy of 94.5% was achieved on visual modality. For bimodality, combining visual and textual modalities produced the best accuracy of 96.21%. The integration of visual, audio, and textual modalities achieved an accuracy of 96.55%.

To learn the intra- and inter-modality dynamics in an end-to-end way, Zadeh et al. [174] proposed a tensor fusion network (TFN). In their network, inter-modality dynamics were modeled by tensor fusion that explicitly aggregates unimodal, bimodal, and trimodal interactions. While intra-modality dynamics were modeled through three modality embedding subnetworks for language, visual and acoustic modalities. In [175], both feature- and decision-level fusion methods were developed to merge emotional information extracted from three modalities. The visual feature set achieved a better precision of 68.1% than the other two modalities. The feature-level fusion of audio and

Table 4 (*continued*).

[25]	Voice active detection, data segment	Audio features extracted by CNN, visual features extracted by 3D CNN	Feature-level: DBN	ANN SVM	RML eNTERFACE05 BAUM-1s	CV LOSO CV LOSGO CV LOSGO	82.38% 85.69% 59.17%
[17]	Key frame selection, data augmentation	Audio features extracted by CNN, visual features extracted by 3D CNN	Feature-level: ELM	ANN SVM	Big data	Train:Val:Test = 70:5:25	99.9%
					eNTERFACE05	CV 5F	86.4%
[165]	Data augmentation	Multi-head attention modules for audio and visual features	Feature-level: Transformer fusion	ANN	AVEC (2017)	$S_{train} = 36$, $S_{dev} = 14$, $S_{test} = 16$	V-A: 65.4–70.8%
[167]	Image graying	HMM for emotion-specific vowels in audio modality and PCA+ANN+GA for visual modality	Decision-level: Weighted summation	GA based ANN	Own	Unknown	97%
[168]	Face detection and alignment	openSMILE for audio features, CNN-RNN and C3D for visual features	Decision-level: weighted summation	SVM ANN	AFEW 6.0	$V_{train} = 774$, $V_{val} = 383$, $V_{test} = 593$	59.02%
[169]	Face detection and alignment	CNN-LSTM for audio features, Semi-supervised Learning with 3D Autoencoder, Convolutional 3D with Auxiliary Network, and CNN-LSTM for visual features	Decision-level: Emotion adaptive fusion	ANN	MMI CASME II	CV LOSO CV LOSO	78.61% 60.98%
[16]	Face detection and alignment, data augmentation	Multitask CNN and openSMILE for audio features extraction, Multitask CNN and LBP for visual features extraction	Decision-level: blending algorithm	SVM	eNTERFACE05	$S_{train} = 30$, $S_{test} = 12$	81.36%
[47]	Face detection and alignment, data augmentation	CNN for audio features extraction, 3D CNN for spatio-temporal facial expression features extraction	Feature-level: cross-attentional fusion	ANN	RECOLA	$S_{train} = 16$, $S_{val} = 15$, $S_{test} = 15$	V-A: 69–83.8%
					fatigue data	Train:Val = 80:20	42.1%

Legenda: V-A: Valence-Arousal; CV: Cross Validation; LOSO: Leave-One-Subject-Out; LOSGO: Leave-One-Speaker-Group-Out;
(N) F: (N) Fold Cross Validation; Train: Training set; Val: Validation set; Test: Testing set; Dev: Development set;
S_*: the number of subjects in different subsets; V_*: the number of videos in different subsets.

visual obtained the best precision of 73.21%. Furthermore, the feature-level fusion of three modalities achieved better precision of 78.2% than decision-level fusion (75.2%). Additionally, through the experiment results, it was found that gaze- and smile-based facial expression features are useful for emotion classification.

To detect and track emotions in conversations, Majumder et al. [41] presented an RNN-based neural architecture named conversational memory networks (CMN). In CMN, textual, audio, and visual features were extracted using CNN, 3D-CNN, and openSMILE. They assumed that the emotion in an utterance is related to three factors: (i) the speaker, (ii) the context given by the preceding utterances, and (iii) the emotion behind the preceding utterances. A party state was used to model each party, and a global state was used to model the context of an utterance. The final emotion classification was realized by emotion representation, which was inferred from the party state of the speaker along with the preceding speakers' states. Jiang et al. [99] proposed a new probability and integrated learning (PIL) based classification algorithm. They presented the topology of integrated learning by simulating the mode and construction of human thinking. Sixteen and fourteen handcrafted visual and audio emotional related features were extracted. Specifically, they extracted lyrics features and split them into positive and negative lyrics. Visual, audio, and lyrics classifiers were trained for emotion recognition. Results from these classifiers were combined through an integration learning technology. The emotion tube can be generated by PIL, which describes the emotional fluctuation.

6.4. Multiple physiological signals fusion

As described in Section 5.4, physiological signals reflect the actual emotional states of human beings because they are rarely controlled subconsciously. Thus, the results of emotion recognition based on physiological signals are more convincing and robust than those based on facial expression, voice, and text. In recent years, multimodal physiological signals fusion methods have been proposed to improve the recognition accuracy further and explore the correlations between EEG and other peripheral physiological signals. For feature-level fusion, Hassan et al. [176] proposed to learn the high-level physiological features from EDA, EMG, and photoplethysmogram (PPG) signals using DBN. Then, the statistical and deep features were fused and fed into the fine Gaussian support vector machine (FGSVM) for emotion recognition. Their method achieved an accuracy of 89.53% on the DEAP dataset.

Zhang et al. [12] proposed a regularized deep fusion framework to fuse representations of various physiological signals. They adopted ensemble deep kernel machine optimization to learn feature representations of different physiological signals. Then, intermediate fusion representations between any two signals were obtained by a one-layer fully connected network. A final representation was generated in a global fusion layer using all representations. Through this operation, the correlation and diversity between different feature representations were well explored, which contributes to improving the quality of global fusion representation. They compared the emotion prediction accuracies of different combinations of physiological signals. Combining all signals, such as EEG, EMG, GSR, and RES, on the DEAP dataset was observed to achieve the best accuracies (64.5% at valence and 63.1% at arousal).

Most fusion methods usually choose the particular classification method, which ignores the different distribution of multiple signals. Zhang et al. [177] proposed to combine EEG, EOG, and EMG signals for emotion prediction. They constructed K-nearest neighbor (KNN), random forest (RF), and CART as base classifiers and ensemble those classifiers by bagging strategy. This method achieved an average accuracy of 94.22% and 90.74% for the two- and four-class tasks. The experimental results also indicated that EEG is the most important signal for emotion prediction. EOG and EMG have a similar distribution. Besides, the ensemble strategy is more effective than using only one classifier because it affects the recognition performance, and it is difficult to select the best classifier.

In [178], a brain-computer music interface system (BCMI) was designed to provide a tool that allows the interaction between an individual undergoing therapy and ongoing music generation. It was concluded that music-induced emotions prove more significant inter-participant differences than emotions aroused by images or videos. They conducted experiments on their collected data. Accuracies of 62.9% and 52.5% were obtained from EEG and peripheral signals, respectively. Furthermore, a consistent result was obtained that EEG performs better than peripheral signals. After fusing these two features, 68.8% accuracy was achieved. The integration of EEG and other physiological features can significantly improve emotion recognition performance, demonstrating that it is crucial to observe brain signals and analyze the peripheral physiological signals associated with emotional responses for emotion understanding.

For decision-level fusion, in [179], EEG, EOG, and GSR signals were combined for multimodal emotion recognition. First, they extracted EEG and peripheral physiological signals (EOG and GSR) features by separately using High Scale and Low Scale CNN models. Emotion classification probabilities of each modality were also calculated. Besides, they calculated the classification score reliability between various labels by Euclidean distance. Finally, emotion recognition results were obtained by combining classification probability and classification score. They compared the accuracy of different EEG frequency bands and peripheral signals on the DEAP dataset. For EEG frequency bands, beta achieved the best accuracy of around 89.7%, and the combination of these bands produced an accuracy of around 95.8%. For peripheral physiological signals, EOG achieved the best accuracy of 85.73%, and the fusion of multiple signals produced an accuracy of around 97.3%. Finally, an accuracy of around 99.17% was obtained by the combination of EEG and peripheral signals. Results indicated that multimodal fusion is better than using EEG or peripheral physiological signals alone.

6.5. Other multimodal fusion

As an effective complement to facial expression, EEG can be detected and exploited to provide implicit emotion states [180]. In [181], facial expression and EEG were combined for emotion recognition. Facial expressions were detected as four basic emotional states (happiness, neutral, sadness, and fear) by a neural network. The EEG signals were detected as four basic emotions as facial expressions and three emotion intensity levels (strong, ordinary, and weak) by two SVMs. Final classification results were obtained by decision-level fusion strategy. In [182], EEG and eye movement were combined to integrate the internal emotional states and external behaviors to enhance the performance of single modalities. The shallow features of EEG and eye movements were fed to a bimodal deep auto-encoder (BDAE) to extract the shared representations. Their experimental results indicated that EEG and eye movement are complementary, and combining these two modalities facilitates the performance improvement of emotion recognition. Specifically, different from the conventional use of 62 electrodes in EEG analysis, they adopted only six symmetrical temporal electrodes to collect EEG signals, which is beneficial for real applications. Some of the recent multimodal emotion recognition methods are listed in Table 5.

Huang et al. [183] developed a deep multimodal attentive fusion (DMAF) model to extract the discriminative features and exploit the complex relationship between visual and text modalities. Feature-level fusion was designed to learn the internal correlation between image and text. The final emotion prediction was obtained through a late fusion scheme combining three attention models. The comprehensive and non-redundant information was exploited from different modalities, and enhanced emotion prediction results were obtained. Many multimodal emotion recognition methods only consider information about the individual who expressed emotions, neglecting contextual information that provides crucial supplementary information. Thus, considering the context information contributes to improving the recognition precision. In [185], the multi-task CNN approach was used to learn body and scene features for context-based emotion recognition simultaneously. Body and scene features were separately extracted by the Xception network and VGG16. Then, the obtained features were concatenated and sent to a fully connected layer for emotional decision. With the fusion of body-related features and context information, the precision of multiple emotions has been improved.

Compared with the emotion recognition of individual utterances, conversational speech emotion analysis has attracted increasing attention due to the vast application in human–computer interaction [189, 190]. Lian et al. [36] proposed a multimodal learning framework using relation and dependencies among utterances for conversational emotion recognition. They proposed an audio-text-speaker fusion component to fuse features from different modalities. The fused segment features were fed into the self-attention based GRU to learn the long-term dependence and contextual information. With the speaker considered, the accuracy improved from 76.4% to 78.02% compared to fusions of audio and text modalities. In their further work [184], a dialogical emotion correction network (DECN) was developed to model human conversation interactions by employing a graphical network. In their work, to automatically correct some errors generated by emotion recognition and further enhance recognition performance, the contextual information and human interactions in conversations were considered.

While the works mentioned above focused on bimodality, the work in [186] proposed a multi-fusion residual memory network to combine facial expression, voice, and text information. They utilized a view-specific learning module (VSL) to explore the intra-modality dynamic of single modalities and used a bidirectional gated recurrent unit network to extract context-dependent representations. A multi-stage fusion module was designed to fuse features of all modalities with a hierarchical fusion strategy to explore time-dependent interactions between modalities. In the first stage, emotion intensity attention was used to capture the critical time steps from acoustic information. In the second stage, time-step level fusion was used to model time-restricted interactions of different modalities and generate fused features for each time step. They used a cross-view learning model to explore inter-modality interactions and fuse features to model the long-term dependency of all time steps in an utterance. Finally, intra-modality and inter-modality information were fused for emotion inference. They compared the performance of single modalities, bimodality, and triple modalities. Results showed that triple modalities fusion facilitates performance improvement because confidential information in different modalities can be integrated.

Ranganathan et al. [51] proposed a convolutional DBN (CDBN) to learn and fuse vocal, facial expression, body gestures, and physiological signal features for the task of multimodal emotion recognition. The complex non-linear feature relationships between the different modalities were explored in an unsupervised manner. The robust multimodal features were generated in an unsupervised manner by the designed DBN model. Besides, CDBN coupled with region of interest extraction learned salient multimodal features, which can enhance the performance of subtle expression recognition. In [188], visual and aural modalities were integrated for enhanced emotion recognition. For visual modality, face, body, and context features were extracted by the pre-trained ResNet50. These features were concatenated and fed into LSTM to learn the temporal dependency of frames in a sequence and produce emotion predictions. For aural modality, Mel-spectrogram

Table 5
Overview of multimodal fusion methods for emotion recognition.

Reference	Modality	Feature extraction	Classifier	Fusion strategy	Dataset	Experimental condition	Accuracy
[183]	V+T	Visual features extracted by CNN, text sentiment features extracted by LSTM	Visual attention model, semantic attention model	Decision-level: weighted combination	Getty Image Twitter Flickr-w Flickr-m	Train:Val:Test = 70:10:20	86.9% 76.3% 85.9% 88%
[180]	V+P	Facial landmark features, EEG, ECG, and GSR features	SVM, Naïve Bayes (NB)	Decision-level: weighted combination	Private data	CV LOSO	V-A-L: 60–59%–58%
[181]	V+P	AdaBoost algorithm for face feature extraction and PSD feature from EEG signal	SVM	Decision-level: sum strategy decision-making strategy	Own	CV	81.25% 82.75%
[182]	V+P	33 eye movement features, and PSD and DE features from EEG	SVM	Feature-level: concatenation	Own	Cross session	72.39%
[36]	A+T	openSMILE for audio feature extraction, deep model for text feature extraction	ANN	Feature-level: GRU	IEMOCAP	$S_{train} = 8$, $S_{test} = 2$	78.02%
[184]	A+T	openSMILE and CNN for audio and text feature extractions, graphical network for context information extraction	SVM, ANN	Feature-level: multi-head attention based GRU	IEMOCAP MELD	$S_{train} = 8$, $S_{test} = 2$ $S_{train} = 260$, $S_{val} = 47$, $S_{test} = 100$	78.08% 56.67%
[185]	V+B	Xception network for body feature extraction and VGG16 for scene feature extraction	ANN	Feature-level: concatenation	EMOTIC	$I_{train} = 12957$, $I_{val} = 3334$, $I_{test} = 7280$	28.33% Macro-precision
[178]	MP	EEG, GSR, BVP, and ECG features	SVM with polynomial kernel	Feature-level: concatenation	Private data	Cross session	V-A: 68.8–68.6%
[176]	MP	9 statistical features of EDA, PPG and zEMG signals, 9 PSD features of EDA, PPG and zEMG signals	Fine Gaussian SVM (FGSVM)	Feature-level: DBN	DEAP	CV 10F	89.53%
[12]	MP	Representation learning layer for EEG, EMG, GSR, RES, MEG, ECG, and EOG feature extractions	ANN	Feature-level: a regularized deep fusion of kernel machines (RDFKMs)	DEAP DECAF	CV LOSO CV LOSO	V-A: 69.6–70.1% V-A: 71.9–60.5%
[177]	MP	Activity, mobility, and complexity features from EEG, EOG, and EMG	KNN, RF, and CART based bagging	Decision-level: ensemble	DEAP	Unknown	V-A: 94.02–94.22%
[179]	MP	Multiscale CNN for EEG, EOG, EMG, and GSR feature extractions	ANN	Decision-level: biologically inspired fusion	DEAP AMIGOS	CV 10F CV 10F	98.52% 99.89%
[172]	A+V+T	Gaussian RBM model used to extract visual and audio features, Replicated Softmax used for textual feature mining	SVM with Gaussian RBF kernel	Feature-level: DBM	YouTube	Unknown	49.9%
[173]	A+V+T	JAudio toolkit for audio feature extraction, Luxand software for facial characteristic points detection, and Bag of concept, sentic, and negation features for text	SVM, ELM	Feature-level: concatenation	eNTERFACE	CV 10F	87.95%

(continued on next page)

representations were fed into ResNet18 model to extract high-level features, which were sent to LSTM for segment-level predictions. Final emotion recognition results were obtained by fusing the results of single modalities at the decision-level fusion. In their study, pose and context information as supplementary streams for face contributes to boosting the performance of visual modality. Results demonstrated the superiority of multimodal fusion to single modalities for emotion recognition. In [187], the context information in images and the interactions among

Table 5 (*continued*).

[19]	A+V+T	openSMILE toolkit for audio feature extraction, convolutional RNN for facial feature learning, and CNN for textual feature	SVM, ANN	Feature-level: concatenation	MOUD IEMOCAP	CV 10F $S_{train} = 8$, $S_{test} = 2$	96.55% 79.35%
[174]	A+V+T	COVAREP acoustic analysis framework for audio feature extraction, DNN for facial feature learning, and LSTM for textual feature learning	ANN	Feature-level: Tensor fusion	CMU-MOSI	CV 5F	Binary: 77.1%
[175]	A+V+T	OpenEAR for audio feature extraction, Luxand software for facial feature extraction, and concepts extraction for text	SVM, ANN, and ELM	Feature-level: concatenation Decision-level: weighted summation	YouTube	CV 10F	78.2% 75.2%
[41]	A+V+T	openSMILE and 3D-CNN for audio and visual feature extractions, CNN for textual feature extraction, GRU for context information	ANN	Feature-level: concatenation	IEMOCAP	$S_{train} = 8$, $S_{test} = 2$	63.4%
[99]	A+V+T	14 audio features, 16 visual features, positive and negative lyrics features	ANN, KNN	Decision-level: linear transformation	Own	Train:Test=3:1	84.3%
[186]	A+V+T	VS-GRU for audio, visual, and textual feature extractions	ANN	Feature-level: Multi-fusion residual memory network	CMU-MOSI CMU-MOSEI IEMOCAP IMDB	$U_{train} = 1284$, $U_{test} = 686$ $U_{train} = 16\,265$, $U_{test} = 4643$ $S_{train} = 8$, $S_{test} = 2$ R_{train}: $R_{test} = 1{:}1$	82.3% 82.4% 83.45% 88.19%
[187]	A+V+T	Self-attention-based CNN for audio, facial, text, and pose feature extraction, context information extraction	ANN	Hybrid fusion: multiplicative fusion	IEMOCAP CMU-MOSEI	Standard training, validation, and testing sets	78.2%
[51]	A+V+P	180 vocal features, 540 facial expression features, 540 body gesture features, and 120 physiological signal features	SVM with radial basis function (RBF) kernels	Feature-level: DBN	emoFBVP CK Mind Reading DEAP MAHNOB-HCI	CV LOSO	83.18% 97.3% 93.4% 79.5% 58.5%
[188]	A+V+B	ResNet50 for face, body, and context feature extractions, ResNet18 for audio feature extraction	ANN	Feature-level: concatenation	Aff-Wild2	Standard training and validation sets	66.8%

Legenda: A: Audio; V: Video; T: Text; P: Physiology; B: Body movement; MP: Multiple physiological signals; V-A-L: Valence-Arousal-Liking; V-A: Valence-Arousal; Train: Training set; Val: Validation set; Test: Testing set; CV: Cross Validation; LOSO: Leave-One-Speaker-Out; (N) F: (N) Fold cross-validation; S_*: the number of subjects in different subsets; I_*: the number of images in different subsets; U_*: the number of utterances in different subsets; R_*: the number of reviews in different subsets.

people were taken as two context channels. With the combination of different modalities (face, audio, text, and pose) and context channels, 78.2% accuracy was achieved on the IEMOCAP dataset.

6.6. Discussion

From all the investigated literature on multimodal emotion recognition, it can be concluded that with the proper fusion strategy, the accuracy of multimodal emotion recognition is outperformed by its unimodal counterparts. Since there are multiple differences among the listed studies, it is improper to directly compare the accuracies across studies. To promise justification, the study-level factors should be held constant [5]. Therefore, the accuracy comparisons of multimodal emotion recognition to single modality emotion recognition are conducted in the same study, instead of across studies. It is calculated that from unimodal to multimodal fusion, the accuracy increases from 2.35% to 19.73%, where the unimodal accuracy is the average of all used modalities. This result indicates that by properly fusing different modalities, emotional recognition performance can be enhanced because different modalities complement each other.

All the proposed methods of multimodal emotion recognition make efforts to improve the performance from different perspectives, for instance, distinct feature representations considering the characteristics of individual modalities, well-designed fusion strategy exploiting the complementarity of multiple modalities, and classifier with the strong capability to distinguish between categories. The advantages of these methods have been highlighted while their own respective limitations in terms of reliability, robustness, and efficiency should not be ignored. For instance, in feature-level fusion, the heterogeneity of different signals is neglected and the reliability of features concatenation needs further discussion. Besides, cross-domain emotion recognition has attracted some attention, while how to deeply explore the potentials of multimodal fusion to simultaneously solve cross-subject, cross-session, and cross-culture still needs much effort.

7. Conclusions and future work

7.1. Conclusions

Multimodal emotion recognition has attracted increasing attention among the emotion recognition community. This paper reviews the recent promising research on multimodal emotion recognition from aspects of multimodal emotion datasets, data preprocessing, unimodal feature extraction, and multimodal information fusion. It can be concluded that, for emotion feature extraction, classical and advanced handcrafted descriptors are still used and developed to extract discriminant emotional features. These feature extraction approaches are essential for the development of emotion recognition since they combine prior knowledge and enable the model with the ability to explain. With the development of DL technology, it has been widely adopted to extract high-level emotion features from different modalities, especially facial images and text.

In addition, fusion strategy plays a crucial role in high-performance multimodal emotion recognition. DL-based algorithms have been developed to fuse features and learn discriminative representation. There have recently been a number of new concepts and strategies aimed at exploring the relationship and complementarity between different modalities and maintaining the unique characteristics of individual modalities. For emotion recognition, traditional machine learning-based classifiers are still the mainstream in different fusion frameworks, such as SVM, RF, ELM, and many others. Also, DL based end-to-end framework has continuously emerged for emotion recognition, which can directly output the classification results to the given input.

Multimodal emotion recognition methods have a wide range of potential applications across various fields, including human–computer interaction (HCI), education, psychology, and neuroscience. In HCI systems, it has the potential to create more natural and effective human–computer interfaces that can understand and respond to users' emotional states, and enhance the emotional expressiveness of virtual agents. In education, multimodal emotion recognition can be used to develop intelligent tutoring systems that can adapt to the emotional states of teachers and students, providing personalized feedback. In psychology and neuroscience, multimodal emotion recognition can provide insights into the neural mechanisms underlying emotional processing and help in the diagnosis and treatment of psychological disorders, such as identifying subtle changes in emotion regulation in patients with mood disorders. By enabling the recognition of emotions across multiple modalities, it has the potential to transform our understanding of how humans communicate and interact.

7.2. Challenging and future work

Although many contributions have been made to endow the development of multimodal emotion recognition, there are still some key issues need to be solved, which are summarized as follows:

- Basic theoretical research. Emotion recognition-related techniques that focus on exploring representative features for accurate emotion classification or prediction have been well developed. In contrast, the relationship between explicit expression and implicit information in affective computing needs further study. Exploring the relationship is significant for understanding various emotional states represented by different signals. Emotion recognition techniques should be integrated with cognitive techniques to bridge the distance between objective emotion recognition and subjective affect cognition of humans.
- Multimodal emotion datasets regarding the spontaneous emotion in the wild. The present multimodal datasets consist mainly of the acted or spontaneous emotions collected in the controlled labs. However, the constraint environment has less noise, and the real implicit emotions may be suppressed. Consequently, the performance is limited when applying the models trained on the desirable datasets to recognize emotions in real scenarios. Therefore, to facilitate the multimodal emotion recognition system used in real situations, more spontaneous emotion data approximate to real life should be collected.
- Feature extraction methods. Various external or internal factors may impede the emotion-related feature extraction. External factors include environmental noises, sensor noises, shooting angles, and cultural differences. Personal character differences are major internal factors. While, current multimodal emotion recognition research barely considers external and internal factors when designing feature extractors. How to eliminate these irrelevant elements and extract essential emotion-related features are still unresolved. It is significant for improving emotion recognition performance to remove these irrelevant variables and find the intrinsic emotion features.
- Multimodal fusion strategy. Recent research shows a definite trend in using deep learning-based technologies for feature-level fusion. However, the specific characteristics of single modalities are ignored in those methods. Both the inter-relationship among modalities and the specific characteristic of a single modality are vital for recognizing emotion. Therefore, how to explore the relationship of different modalities and retain the particular property of a single modality is waiting to be settled. Additionally, exploiting and analyzing the correlation and characteristics of various modalities contributes to the design effective fusion strategy for multimodal emotion recognition.
- Multimodal emotion recognition in conversation. The conversation takes up most of our lives and work. However, few researchers drew attention to recognizing emotions in conversations. Emotions expressed by one person are often caused or affected by the content or other persons in a conversation, which is difficult to analyze quantitatively. It is challenging to recognize emotions in the conversation because various elements should be considered, such as individual emotion, context, and the influence of others. If the emotion recognition in conversations is well solved, that will facilitate the advancement and application of multimodal emotion recognition.

CRediT authorship contribution statement

Bei Pan: Investigation, Writing – original draft. **Kaoru Hirota:** Conceptualization, Supervision. **Zhiyang Jia:** Conceptualization, Writing – review & editing. **Yaping Dai:** Resources, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the China Scholarship Council under Grant No. 202206030126, National Talents Foundation under Grant No. WQ20141100198 and the Beijing Municipal Natural Science Foundation under Grant No. L191020.

References

[1] R.W. Picard, Affective Computing, MIT Press, 2000.

[2] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A survey of affect recognition methods: Audio, visual, and spontaneous expressions, IEEE Trans. Pattern Anal. Mach. Intell. 31 (1) (2008) 39–58.

[3] R.A. Calvo, S. D'Mello, Affect detection: An interdisciplinary review of models, methods, and their applications, IEEE Trans. Affect. Comput. 1 (1) (2010) 18–37.

[4] C.-H. Wu, J.-C. Lin, W.-L. Wei, Survey on audiovisual emotion recognition: Databases, features, and data fusion strategies, APSIPA Trans. Signal Inf. Process. 3 (2014).

[5] S.K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Comput. Surv. 47 (3) (2015) 1–36.

[6] S. Zhao, S. Wang, M. Soleymani, D. Joshi, Q. Ji, Affective computing for large-scale heterogeneous multimedia data: A survey, ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 15 (3s) (2019) 1–32.

[7] Y. Jiang, W. Li, M.S. Hossain, M. Chen, A. Alelaiwi, M. Al-Hammadi, A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition, Inf. Fusion 53 (2020) 209–221.

[8] N.J. Shoumy, L.-M. Ang, K.P. Seng, D.M. Rahaman, T. Zia, Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals, J. Netw. Comput. Appl. 149 (2020) 102447.

[9] P. Ekman, The argument and evidence about universals in facial expressions, in: Handbook of Social Psychophysiology, Vol. 143, Wiley, Chichester, England, 1989, p. 164.

[10] J.A. Russell, Affective space is bipolar, J. Personal. Soc. Psychol. 37 (3) (1979) 345.

[11] A. Mehrabian, Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament, Curr. Psychol. 14 (4) (1996) 261–292.

[12] J. Zhang, Z. Yin, P. Chen, S. Nichele, Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review, Inf. Fusion 59 (2020) 103–126.

[13] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE'05 audio-visual emotion database, in: 22nd International Conference on Data Engineering Workshops (ICDEW'06), IEEE, 2006, p. 8.

[14] E. Ghaleb, M. Popa, S. Asteriadis, Metric learning-based multimodal audio-visual emotion recognition, IEEE Multimedia 27 (1) (2019) 37–48.

[15] S. Pini, O.B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, B. Huet, Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, 2017, pp. 536–543.

[16] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, P. Xiao, Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features, Neurocomputing 391 (2020) 42–51.

[17] M.S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio-visual emotional big data, Inf. Fusion 49 (2019) 69–78.

[18] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, G. Anbarjafari, Audio-visual emotion recognition in video clips, IEEE Trans. Affect. Comput. 10 (1) (2017) 60–75.

[19] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional MKL based multimodal emotion recognition and sentiment analysis, in: 2016 IEEE 16th International Conference on Data Mining (ICDM), IEEE, 2016, pp. 439–448.

[20] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, C. Fookes, Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition, Comput. Vis. Image Underst. 174 (2018) 33–42.

[21] M. Bejani, D. Gharavian, N.M. Charkari, Audiovisual emotion recognition using ANOVA feature selection method and multi-classifier neural networks, Neural Comput. Appl. 24 (2) (2014) 399–412.

[22] S. Dobrišek, R. Gajšek, F. Mihelič, N. Pavešić, V. Štruc, Towards efficient multi-modal emotion recognition, Int. J. Adv. Robot. Syst. 10 (1) (2013) 53.

[23] Y. Wang, L. Guan, Recognizing human emotional state from audiovisual signals, IEEE Trans. Multimed. 10 (5) (2008) 936–946.

[24] S. Zhang, S. Zhang, T. Huang, W. Gao, Q. Tian, Learning affective features with a hybrid deep model for audio-visual emotion recognition, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 3030–3043.

[25] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, A. Košir, Audio-visual emotion fusion (AVEF): A deep efficient weighted approach, Inf. Fusion 46 (2019) 184–192.

[26] I. Kansizoglou, L. Bampis, A. Gasteratos, An active learning paradigm for online audio-visual emotion recognition, IEEE Trans. Affect. Comput. (2019) 1.

[27] R.R. Sarvestani, R. Boostani, FF-SKPCCA: Kernel probabilistic canonical correlation analysis, Appl. Intell. 46 (2) (2017) 438–454.

[28] N.E.D. Elmadany, Y. He, L. Guan, Multiview emotion recognition via multi-set locality preserving canonical correlation analysis, in: 2016 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2016, pp. 590–593.

[29] S. Zhalehpour, O. Onder, Z. Akhtar, C.E. Erdem, BAUM-1: A spontaneous audio-visual face database of affective and mental states, IEEE Trans. Affect. Comput. 8 (3) (2016) 300–313.

[30] X. Pan, S. Zhang, W. Guo, X. Zhao, Y. Chuang, Y. Chen, H. Zhang, Video-based facial expression recognition using deep temporal-spatial networks, IETE Tech. Rev. 37 (4) (2020) 402–409.

[31] L. Singh, S. Singh, N. Aggarwal, Improved TOPSIS method for peak frame selection in audio-video human emotion recognition, Multimedia Tools Appl. 78 (5) (2019) 6277–6308.

[32] J. Cornejo, H. Pedrini, Bimodal emotion recognition based on audio and facial parts using deep convolutional neural networks, in: 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2019, pp. 111–117.

[33] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: A multimodal multi-party dataset for emotion recognition in conversations, 2018, arXiv preprint arXiv:1810.02508.

[34] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, G. Zhou, Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2019, pp. 5415–5421.

[35] Y. Zhang, Q. Li, D. Song, P. Zhang, P. Wang, Quantum-inspired interactive networks for conversational sentiment analysis, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI), 2019.

[36] Z. Lian, J. Tao, B. Liu, J. Huang, Conversational emotion analysis via attention mechanisms, 2019, arXiv preprint arXiv:1910.11263.

[37] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, A. Gelbukh, Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, 2019, arXiv preprint arXiv:1908.11540.

[38] P. Zhong, D. Wang, C. Miao, Knowledge-enriched transformer for emotion detection in textual conversations, 2019, arXiv preprint arXiv:1909.10681.

[39] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, S.S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Lang. Resour. Eval. 42 (4) (2008) 335–359.

[40] D. Hazarika, S. Poria, R. Zimmermann, R. Mihalcea, Emotion recognition in conversations with transfer learning from generative conversation modeling, 2019, arXiv preprint arXiv:1910.04980.

[41] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, E. Cambria, DialogueRNN: An attentive RNN for emotion detection in conversations, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 6818–6825.

[42] S. Poria, E. Cambria, A.A.D. Hazarika, N. Majumder, A. Zadeh, L.-P. Morency, Context-dependent sentiment analysis in user-generated videos, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 873–883.

[43] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, R. Zimmermann, Conversational memory network for emotion recognition in dyadic dialogue videos, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2018, NIH Public Access, 2018, pp. 2122–2132.

[44] G. McKeown, M. Valstar, R. Cowie, M. Pantic, M. Schroder, The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent, IEEE Trans. Affect. Comput. 3 (1) (2011) 5–17.

[45] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, R. Zimmermann, ICON: Interactive conversational memory network for multimodal emotion detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2594–2604.

[46] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–8.

[47] R. Gnana Praveen, G. Eric, C. Patrick, Cross attentional audio-visual fusion for dimensional emotion recognition, 2021, arXiv preprint arXiv:2111.05222.

[48] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M.A. Nicolaou, B. Schuller, S. Zafeiriou, Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 5200–5204.

[49] P. Tzirakis, G. Trigeorgis, M.A. Nicolaou, B.W. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, IEEE J. Sel. Top. Sign. Proces. 11 (8) (2017) 1301–1309.

[50] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, A multimodal database for affect recognition and implicit tagging, IEEE Trans. Affect. Comput. 3 (1) (2011) 42–55.

[51] H. Ranganathan, S. Chakraborty, S. Panchanathan, Multimodal emotion recognition using deep learning architectures, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2016, pp. 1–9.

[52] B. Nakisa, M.N. Rastgoo, D. Tjondronegoro, V. Chandran, Evolutionary computation algorithms for feature selection of EEG-based emotion recognition using mobile sensors, Expert Syst. Appl. 93 (2018) 143–155.

[53] P. Li, H. Liu, Y. Si, C. Li, F. Li, X. Zhu, X. Huang, Y. Zeng, D. Yao, Y. Zhang, et al., EEG based emotion recognition by combining functional connectivity network and local activations, IEEE Trans. Biomed. Eng. 66 (10) (2019) 2869–2881.

[54] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, DEAP: A database for emotion analysis using physiological signals, IEEE Trans. Affect. Comput. 3 (1) (2011) 18–31.

[55] Z. Yin, M. Zhao, Y. Wang, J. Yang, J. Zhang, Recognition of emotions using multimodal physiological signals and an ensemble deep learning model, Comput. Methods Programs Biomed. 140 (2017) 93–110.

[56] F. Ren, Y. Dong, W. Wang, Emotion recognition based on physiological signals using brain asymmetry index and echo state network, Neural Comput. Appl. 31 (9) (2019) 4491–4501.

[57] Y. Liu, Y. Ding, C. Li, J. Cheng, R. Song, F. Wan, X. Chen, Multi-channel EEG-based emotion recognition via a multi-level features guided capsule network, Comput. Biol. Med. 123 (2020) 103927.

[58] J. Ma, H. Tang, W.-L. Zheng, B.-L. Lu, Emotion recognition using multi-modal residual LSTM network, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 176–183.

[59] A. Kumar, A. Kaur, M. Kumar, Face detection techniques: A review, Artif. Intell. Rev. 52 (2) (2019) 927–948.

[60] P. Viola, M. Jones, 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, Vol. 1, IEEE, 2001, p. 1.

[61] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5325–5334.

[62] G. Guo, H. Wang, Y. Yan, J. Zheng, B. Li, A fast face detection method via convolutional neural network, Neurocomputing 395 (2020) 128–137.

[63] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.

[64] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.

[65] A.T. Lopes, E. De Aguiar, A.F. De Souza, T. Oliveira-Santos, Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order, Pattern Recognit. 61 (2017) 610–628.

[66] S. Shan, W. Gao, B. Cao, D. Zhao, Illumination normalization for robust face recognition against varying lighting conditions, in: 2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443), IEEE, 2003, pp. 157–164.

[67] A.V. Oppenheim, R.W. Schafer, From frequency to quefrency: A history of the cepstrum, IEEE Signal Process. Mag. 21 (5) (2004) 95–106.

[68] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: Global and local perception GAN for photorealistic and identity preserving frontal view synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.

[69] J. Pohjalainen, F. Fabien Ringeval, Z. Zhang, B. Schuller, Spectral and cepstral audio noise reduction techniques in speech emotion recognition, in: Proceedings of the 24th ACM International Conference on Multimedia, 2016, pp. 670–674.

[70] M.A. Palomino, F. Aider, Evaluating the effectiveness of text pre-processing in sentiment analysis, Appl. Sci. 12 (17) (2022) 8765.

[71] E.H. Houssein, A. Hammad, A.A. Ali, Human emotion recognition from EEG-based brain–computer interface using machine learning: A comprehensive review, Neural Comput. Appl. (2022) 1–31.

[72] B. Fasel, J. Luettin, Automatic facial expression analysis: A survey, Pattern Recognit. 36 (1) (2003) 259–275.

[73] C.A. Corneanu, M.O. Simón, J.F. Cohn, S.E. Guerrero, Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications, IEEE Trans. Pattern Anal. Mach. Intell. 38 (8) (2016) 1548–1568.

[74] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, U. Schmid, Automatic detection of pain from facial expressions: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 43 (6) (2019) 1815–1831.

[75] S. Li, W. Deng, Deep facial expression recognition: A survey, IEEE Trans. Affect. Comput. (2020) 1.

[76] G.R. Alexandre, J.M. Soares, G.A.P. Thé, Systematic review of 3D facial expression recognition methods, Pattern Recognit. 100 (2020) 107108.

[77] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, Y.-J. Liu, Video-based facial micro-expression analysis: A survey of datasets, features and algorithms, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1.

[78] R.A. Calvo, S. D'Mello, J.M. Gratch, A. Kappas, The Oxford Handbook of Affective Computing, Oxford Library of Psychology, 2015.

[79] Y.-L. Tian, T. Kanade, J.F. Cohn, Facial expression analysis, in: Handbook of Face Recognition, Springer, 2005, pp. 247–275.

[80] A. Majumder, L. Behera, V.K. Subramanian, Emotion recognition from geometric facial features using self-organizing map, Pattern Recognit. 47 (3) (2014) 1282–1293.

[81] B. Ryu, A.R. Rivera, J. Kim, O. Chae, Local directional ternary pattern for facial expression recognition, IEEE Trans. Image Process. 26 (12) (2017) 6006–6018.

[82] B. Pan, K. Hirota, Z. Jia, L. Zhao, X. Jin, Y. Dai, Multimodal emotion recognition based on feature selection and extreme learning machine in video clips, J. Ambient Intell. Humaniz. Comput. (2021) 1–15.

[83] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Trans. Pattern Anal. Mach. Intell. 24 (7) (2002) 971–987.

[84] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, IEEE, 2005, pp. 886–893.

[85] L. Zhang, D. Tjondronegoro, Facial expression recognition using facial movement features, IEEE Trans. Affect. Comput. 2 (4) (2011) 219–229.

[86] T. Yun, L. Guan, Human emotional state recognition using real 3D visual features from Gabor library, Pattern Recognit. 46 (2) (2013) 529–538.

[87] Y. Yacoob, L.S. Davis, Recognizing human facial expressions from long image sequences using optical flow, IEEE Trans. Pattern Anal. Mach. Intell. 18 (6) (1996) 636–642.

[88] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1940–1954.

[89] G. Zhao, M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Trans. Pattern Anal. Mach. Intell. 29 (6) (2007) 915–928.

[90] J. Chen, Z. Chen, Z. Chi, H. Fu, Facial expression recognition in video with multiple feature fusion, IEEE Trans. Affect. Comput. 9 (1) (2016) 38–50.

[91] H. Fang, N. Mac Parthaláin, A.J. Aubrey, G.K. Tam, R. Borgo, P.L. Rosin, P.W. Grant, D. Marshall, M. Chen, Facial expression recognition in dynamic sequences: An integrated approach, Pattern Recognit. 47 (3) (2014) 1271–1281.

[92] E. Sariyanidi, H. Gunes, A. Cavallaro, Learning bases of activity for facial expression recognition, IEEE Trans. Image Process. 26 (4) (2017) 1965–1978.

[93] N. Perveen, D. Roy, K.M. Chalavadi, Facial expression recognition in videos using dynamic kernels, IEEE Trans. Image Process. 29 (2020) 8316–8325.

[94] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[95] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016.

[96] A. Majumder, L. Behera, V.K. Subramanian, Automatic facial expression recognition system using deep network-based data fusion, IEEE Trans. Cybern. 48 (1) (2016) 103–114.

[97] M. Verma, S.K. Vipparthi, G. Singh, S. Murala, LEARNet: Dynamic imaging network for micro expression recognition, IEEE Trans. Image Process. 29 (2019) 1618–1627.

[98] M. Wu, W. Su, L. Chen, Z. Liu, W. Cao, K. Hirota, Weight-adapted convolution neural network for facial expression recognition in human-robot interaction, IEEE Trans. Syst. Man Cybern.: Syst. (2019) 1473–1484.

[99] D. Jiang, K. Wu, D. Chen, G. Tu, T. Zhou, A. Garg, L. Gao, A probability and integrated learning based classification algorithm for high-level human emotion recognition problems, Measurement 150 (2020) 107049.

[100] K. Wang, X. Peng, J. Yang, S. Lu, Y. Qiao, Suppressing uncertainties for large-scale facial expression recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6897–6906.

[101] Y. Fu, X. Wu, X. Li, Z. Pan, D. Luo, Semantic neighborhood-aware deep facial expression recognition, IEEE Trans. Image Process. 29 (2020) 6535–6548.

[102] H.-W. Ng, V.D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015, pp. 443–449.

[103] G. Pons, D. Masip, Multitask, multilabel, and multidomain learning with convolutional networks for emotion recognition, IEEE Trans. Cybern. (2020) 1–8.

[104] F. Zhang, T. Zhang, Q. Mao, C. Xu, Joint pose and expression modeling for facial expression recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3359–3368.

[105] F. Zhang, T. Zhang, Q. Mao, C. Xu, Geometry guided pose-invariant facial expression recognition, IEEE Trans. Image Process. 29 (2020) 4445–4460.

[106] Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using CNN with attention mechanism, IEEE Trans. Image Process. 28 (5) (2018) 2439–2450.

[107] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, Z. Luo, Conditional convolution neural network enhanced random forest for facial expression recognition, Pattern Recognit. 84 (2018) 251–261.

[108] J. Lee, S. Kim, S. Kim, K. Sohn, Multi-modal recurrent attention networks for facial expression recognition, IEEE Trans. Image Process. 29 (2020) 6977–6991.

[109] W. Chen, D. Zhang, M. Li, D.-J. Lee, STCAM: Spatial-temporal and channel attention module for dynamic facial expression recognition, IEEE Trans. Affect. Comput. (2020) 1.

[110] T. Zhang, W. Zheng, Z. Cui, Y. Zong, Y. Li, Spatial–temporal recurrent neural network for emotion recognition, IEEE Trans. Cybern. 49 (3) (2018) 839–847.

[111] W.J. Baddar, S. Lee, Y.M. Ro, On-the-fly facial expression prediction using LSTM encoded appearance-suppressed dynamics, IEEE Trans. Affect. Comput. (2019) 1.

[112] D. Meng, X. Peng, K. Wang, Y. Qiao, Frame attention networks for facial expression recognition in videos, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 3866–3870.

[113] Y. Zong, W. Zheng, Z. Cui, G. Zhao, B. Hu, Toward bridging microexpressions from different domains, IEEE Trans. Cybern. 50 (12) (2019) 5047–5060.

[114] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, G. Zhao, Revealing the invisible with model and data shrinking for composite-database micro-expression recognition, IEEE Trans. Image Process. 29 (2020) 8590–8605.

[115] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, L. Lin, Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1.

[116] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognit. 44 (3) (2011) 572–587.

[117] S.G. Koolagudi, K.S. Rao, Emotion recognition from speech: A review, Int. J. Speech Technol. 15 (2) (2012) 99–117.

[118] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, B.W. Schuller, Multi-task semi-supervised adversarial autoencoding for speech emotion recognition, IEEE Trans. Affect. Comput. (2020) 1.

[119] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, B.W. Schuller, Survey of deep representation learning for speech emotion recognition, IEEE Trans. Affect. Comput. (2021) 1.

[120] S.P. Yadav, S. Zaidi, A. Mishra, V. Yadav, Survey on machine learning in speech emotion recognition and vision systems using a recurrent neural network (RNN), Arch. Comput. Methods Eng. (2021) 1–18.

[121] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Trans. Multimed. 20 (6) (2017) 1576–1590.

[122] D. Issa, M.F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomed. Signal Process. Control 59 (2020) 101894.

[123] K. Ito, T. Fujioka, Q. Sun, K. Nagamatsu, Audio-visual speech emotion recognition by disentangling emotion and identity attributes, in: Proceedings of Interspeech 2021, 2021, pp. 4493–4497.

[124] E. Kalhor, B. Bakhtiari, Speaker independent feature selection for speech emotion recognition: A multi-task approach, Multimedia Tools Appl. 80 (6) (2021) 8127–8146.

[125] L. Chen, W. Su, Y. Feng, M. Wu, J. She, K. Hirota, Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction, Inform. Sci. 509 (2020) 150–163.

[126] C. Yogesh, M. Hariharan, R. Ngadiran, A.H. Adom, S. Yaacob, C. Berkai, K. Polat, A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal, Expert Syst. Appl. 69 (2017) 149–158.

[127] M. Sarma, P. Ghahremani, D. Povey, N.K. Goel, K.K. Sarma, N. Dehak, Emotion identification from raw speech signals using DNNs, in: Interspeech, 2018, pp. 3097–3101.

[128] S. Zhang, X. Zhao, Q. Tian, Spontaneous speech emotion recognition using multiscale deep convolutional LSTM, IEEE Trans. Affect. Comput. (2019) 1.

[129] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, X. Xu, Spatiotemporal and frequential cascaded attention networks for speech emotion recognition, Neurocomputing 448 (2021) 238–248.

[130] L. Guo, L. Wang, J. Dang, E.S. Chng, S. Nakagawa, Learning affective representations based on magnitude and dynamic relative phase information for speech emotion recognition, Speech Commun. 136 (2022) 118–127.

[131] D. Li, J. Liu, Z. Yang, L. Sun, Z. Wang, Speech emotion recognition using recurrent neural networks with directional self-attention, Expert Syst. Appl. 173 (2021) 114683.

[132] Y. Zong, W. Zheng, T. Zhang, X. Huang, Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression, IEEE Signal Process. Lett. 23 (5) (2016) 585–589.

[133] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, G. Hofer, Analysis of deep learning architectures for cross-corpus speech emotion recognition, in: Proceedings of Interspeech 2019, 2019, pp. 1656–1660.

[134] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps, Cross corpus speech emotion classification-an effective transfer learning technique, 2018, arXiv preprint arXiv:1801.06353.

[135] X. Li, M. Akagi, Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model, Speech Commun. 110 (2019) 1–12.

[136] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Universum autoencoder-based domain adaptation for speech emotion recognition, IEEE Signal Process. Lett. 24 (4) (2017) 500–504.

[137] M. Abdelwahab, C. Busso, Domain adversarial for acoustic emotion recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 26 (12) (2018) 2423–2435.

[138] J. Gideon, M. McInnis, E.M. Provost, Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG), IEEE Trans. Affect. Comput. 12 (4) (2019) 1055–1068.

[139] P. Song, W. Zheng, Feature selection based transfer subspace learning for speech emotion recognition, IEEE Trans. Affect. Comput. 11 (3) (2018) 373–382.

[140] P. Song, Transfer linear subspace learning for cross-corpus speech emotion recognition, IEEE Trans. Affect. Comput. 10 (2) (2019) 265–275.

[141] C.O. Alm, D. Roth, R. Sproat, Emotions from text: machine learning for text-based emotion prediction, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 579–586.

[142] Q. Jin, C. Li, S. Chen, H. Wu, Speech emotion recognition with acoustic and lexical features, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2015, pp. 4749–4753.

[143] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, LSTM-based text emotion recognition using semantic and emotional word vectors, in: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), IEEE, 2018, pp. 1–6.

[144] B. Liu, Text sentiment analysis based on CBOW model and deep learning in big data environment, J. Ambient Intell. Humaniz. Comput. 11 (2) (2020) 451–458.

[145] J. Wang, L.-C. Yu, K.R. Lai, X. Zhang, Tree-structured regional CNN-LSTM model for dimensional sentiment analysis, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2019) 581–591.

[146] F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, S. Qiao, Attention-emotion-enhanced convolutional LSTM for sentiment analysis, IEEE Trans. Neural Netw. Learn. Syst. (2021).

[147] M. Egger, M. Ley, S. Hanke, Emotion recognition from physiological signal analysis: A review, Electron. Notes Theor. Comput. Sci. 343 (2019) 35–55.

[148] R. Jenke, A. Peer, M. Buss, Feature extraction and selection for emotion recognition from EEG, IEEE Trans. Affect. Comput. 5 (3) (2014) 327–339.

[149] M. Moghimi, R. Stone, P. Rotshtein, Affective recognition in dynamic and interactive virtual environments, IEEE Trans. Affect. Comput. 11 (1) (2017) 45–62.

[150] S.K. Khare, V. Bajaj, Time-frequency representation and convolutional neural network-based emotion recognition, IEEE Trans. Neural Netw. Learn. Syst. 32 (7) (2020) 2901–2909.

[151] J. Hu, C. Wang, Q. Jia, Q. Bu, R. Sutcliffe, J. Feng, ScalingNet: extracting features from raw EEG data for emotion recognition, Neurocomputing 463 (2021) 177–184.

[152] D. Huang, S. Chen, C. Liu, L. Zheng, Z. Tian, D. Jiang, Differences first in asymmetric brain: A bi-hemisphere discrepancy convolutional neural network for EEG emotion recognition, Neurocomputing 448 (2021) 140–151.

[153] T. Song, W. Zheng, P. Song, Z. Cui, EEG emotion recognition using dynamical graph convolutional neural networks, IEEE Trans. Affect. Comput. 11 (3) (2018) 532–541.

[154] T. Song, W. Zheng, S. Liu, Y. Zong, Z. Cui, Y. Li, Graph-embedded convolutional neural network for image-based EEG emotion recognition, IEEE Trans. Emerg. Top. Comput. (2021).

[155] W. Zheng, Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis, IEEE Trans. Cogn. Dev. Syst. 9 (3) (2016) 281–290.

[156] W.-L. Zheng, B.-L. Lu, Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks, IEEE Trans. Auton. Ment. Dev. 7 (3) (2015) 162–175.

[157] X. Du, C. Ma, G. Zhang, J. Li, Y.-K. Lai, G. Zhao, X. Deng, Y.-J. Liu, H. Wang, An efficient LSTM network for emotion recognition from multichannel EEG signals, IEEE Trans. Affect. Comput. (2020).

[158] P. Zhong, D. Wang, C. Miao, EEG-based emotion recognition using regularized graph neural networks, IEEE Trans. Affect. Comput. (2020) 1.

[159] J. Li, S. Qiu, Y.-Y. Shen, C.-L. Liu, H. He, Multisource transfer learning for cross-subject EEG emotion recognition, IEEE Trans. Cybern. 50 (7) (2019) 3281–3293.

[160] Y. Li, B. Fu, F. Li, G. Shi, W. Zheng, A novel transferability attention neural network model for EEG emotion recognition, Neurocomputing 447 (2021) 92–101.

[161] J. Zhang, L. Xing, Z. Tan, H. Wang, K. Wang, Multi-head attention fusion networks for multi-modal speech emotion recognition, Comput. Ind. Eng. 168 (2022) 108078.

[162] Y. Wei, D. Hu, Y. Tian, X. Li, Learning in audio-visual context: A review, analysis, and new perspective, 2022, arXiv preprint arXiv:2208.09579.

[163] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, L. Huang, What makes multi-modal learning better than single (provably), Adv. Neural Inf. Process. Syst. 34 (2021) 10944–10956.

[164] A.I. Middya, B. Nag, S. Roy, Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities, Knowl.-Based Syst. 244 (2022) 108580.

[165] J. Huang, J. Tao, B. Liu, Z. Lian, M. Niu, Multimodal transformer fusion for continuous emotion recognition, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 3507–3511.

[166] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, Multi-cue fusion for emotion recognition in the wild, Neurocomputing 309 (2018) 27–35.

[167] L.-A. Perez-Gaspar, S.-O. Caballero-Morales, F. Trujillo-Romero, Multimodal emotion recognition with evolutionary computation for human-robot interaction, Expert Syst. Appl. 66 (2016) 42–61.

[168] Y. Fan, X. Lu, D. Li, Y. Liu, Video-based emotion recognition using CNN-RNN and C3D hybrid networks, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, 2016, pp. 445–450.

[169] D.H. Kim, W.J. Baddar, J. Jang, Y.M. Ro, Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition, IEEE Trans. Affect. Comput. 10 (2) (2017) 223–236.

[170] Z. Farhoudi, S. Setayeshi, Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition, Speech Commun. 127 (2021) 92–103.

[171] S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R.C. Ferrari, et al., Combining modality specific deep neural networks for emotion recognition in video, in: Proceedings of the 15th ACM on International Conference on Multimodal Interaction, 2013, pp. 543–550.

[172] L. Pang, C.-W. Ngo, Mutlimodal learning with deep boltzmann machine for emotion prediction in user generated videos, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, 2015, pp. 619–622.

[173] S. Poria, E. Cambria, A. Hussain, G.-B. Huang, Towards an intelligent framework for multimodal affective data analysis, Neural Netw. 63 (2015) 104–116.

[174] A. Zadeh, M. Chen, S. Poria, E. Cambria, L.-P. Morency, Tensor fusion network for multimodal sentiment analysis, 2017, arXiv preprint arXiv:1707.07250.

[175] S. Poria, E. Cambria, N. Howard, G.-B. Huang, A. Hussain, Fusing audio, visual and textual clues for sentiment analysis from multimodal content, Neurocomputing 174 (2016) 50–59.

[176] M.M. Hassan, M.G.R. Alam, M.Z. Uddin, S. Huda, A. Almogren, G. Fortino, Human emotion recognition using deep belief network architecture, Inf. Fusion 51 (2019) 10–18.

[177] J. Zhang, Y. Zhang, S. Zhan, C. Cheng, Ensemble emotion recognizing with multiple modal physiological signals, 2020, arXiv preprint arXiv:2001.00191.

[178] I. Daly, D. Williams, A. Malik, J. Weaver, A. Kirke, F. Hwang, E. Miranda, S.J. Nasuto, Personalised, multi-modal, affective state detection for hybrid brain-computer music interfacing, IEEE Trans. Affect. Comput. 11 (1) (2018) 111–124.

[179] Y. Zhao, X. Cao, J. Lin, D. Yu, X. Cao, Multimodal affective states recognition based on multiscale CNNs and biologically inspired decision fusion model, IEEE Trans. Affect. Comput. (2021).

[180] R. Gupta, M. Khomami Abadi, J.A. Cárdenes Cabré, F. Morreale, T.H. Falk, N. Sebe, A quality adaptive multimodal affect recognition system for user-centric multimedia indexing, in: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, 2016, pp. 317–320.

[181] Y. Huang, J. Yang, P. Liao, J. Pan, Fusion of facial expressions and EEG for multimodal emotion recognition, Comput. Intell. Neurosci. 2017 (2017).

[182] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, A. Cichocki, EmotionMeter: A multimodal framework for recognizing human emotions, IEEE Trans. Cybern. 49 (3) (2018) 1110–1122.

[183] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, Image–text sentiment analysis via deep multimodal attentive fusion, Knowl.-Based Syst. 167 (2019) 26–37.

[184] Z. Lian, B. Liu, J. Tao, DECN: Dialogical emotion correction network for conversational emotion recognition, Neurocomputing 454 (2021) 483–495.

[185] I. Bendjoudi, F. Vanderhaegen, D. Hamad, F. Dornaika, Multi-label, multi-task CNN approach for context-based emotion recognition, Inf. Fusion 76 (2021) 422–428.

[186] S. Mai, H. Hu, J. Xu, S. Xing, Multi-fusion residual memory network for multimodal human sentiment comprehension, IEEE Trans. Affect. Comput. (2020).

[187] T. Mittal, A. Bera, D. Manocha, Multimodal and context-aware Emotion perception model with multiplicative fusion, IEEE MultiMedia 28 (2) (2021) 67–75.

[188] P. Antoniadis, I. Pikoulis, P.P. Filntisis, P. Maragos, An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3645–3651.

[189] S. Poria, N. Majumder, R. Mihalcea, E. Hovy, Emotion recognition in conversation: Research challenges, datasets, and recent advances, IEEE Access 7 (2019) 100943–100953.

[190] Y. Wang, J. Zhang, J. Ma, S. Wang, J. Xiao, Contextualized emotion recognition in conversation as sequence tagging, in: Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2020, pp. 186–195.

Bei Pan received the B.S. degree in Automation from Beijing University of Civil Engineering and Architecture, Beijing, China, in 2016 and the M.S. degree in Testing and Measurement Technology and Instrument from Kunming University of Science and Technology, Kunming, China, in 2019. She is currently working toward the Ph.D. degree in control science and engineering at Beijing Institute of Technology, Beijing, China. Her research interests include emotion recognition and computer vision.

Kaoru Hirota received the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1979. He is currently a professor with the school of automation, Beijing Institute of Technology, Beijing, China. He is also a Professor Emeritus with the Tokyo Institute of Technology, Tokyo, Japan. His current research interests include fuzzy systems, intelligent robot, and image understanding. Dr. Hirota is an Experienced President, a Fellow of the International Fuzzy Systems Association, and a President of the Japan Society for Fuzzy Theory and Systems.

Zhiyang Jia received the B.E. degree from the Northwestern Polytechnical University, Xi'an, China, in 2010, the M.E. degree from the Engineering Center for Digital Community of Ministry of Education, Department of Control Science and Engineering, Beijing University of Technology, Beijing, China, in 2013, and the Ph.D. degree in Electrical and Computer Engineering from the University of Connecticut, Storrs, CT, USA, in 2017. He is currently an Assistant Professor at the School of Automation, Beijing Institute of Technology, Beijing, China. His research interests include smart manufacturing, modeling, analysis, and control of production systems.

Yaping Dai received the B.E. degree in measurement and control technology and instruments from Nanjing Normal University, Nanjing, China, in 1983, the M.E. degree from the School of Automation, Beijing University of Chemical Technology, Beijing, China, in 1990, and the Ph.D. degree from the School of Automation, Beijing Institute of Technology, Beijing, in 1993. She is currently a Professor with the School of Automation, Beijing Institute of Technology. Her current research interests include computational intelligence, fuzzy modeling, knowledge discovery, knowledge-based decision making, pattern recognition, and image processing.