Information Fusion 53 (2020) 209–221

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition

Yingying Jiang[a], Wei Li[a], M. Shamim Hossain[b,*], Min Chen[a,*], Abdulhameed Alelaiwi[b], Muneer Al-Hammadi[c]

[a] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[b] Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[c] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

ARTICLE INFO

Keywords:
Artificial intelligence
Multimodal information fusion
Data-driven emotion recognition

ABSTRACT

With the rapid development of artificial intelligence and mobile Internet, the new requirements for human-computer interaction have been put forward. The personalized emotional interaction service is a new trend in the human-computer interaction field. As a basis of emotional interaction, emotion recognition has also introduced many new advances with the development of artificial intelligence. The current research on emotion recognition mostly focuses on single-modal recognition such as expression recognition, speech recognition, limb recognition, and physiological signal recognition. However, the lack of the single-modal emotional information and vulnerability to various external factors lead to lower accuracy of emotion recognition. Therefore, multimodal information fusion for data-driven emotion recognition has been attracting the attention of researchers in the affective computing filed. This paper reviews the development background and hot spots of the data-driven multimodal emotion information fusion. Considering the real-time mental health monitoring system, the current development of multimodal emotion data sets, the multimodal features extraction, including the EEG, speech, expression, text features, and multimodal fusion strategies and recognition methods are discussed and summarized in detail. The main objective of this work is to present a clear explanation of the scientific problems and future research directions in the multimodal information fusion for data-driven emotion recognition field.

1. Introduction

With the rapid development of mobile Internet and artificial intelligence technology, the more and more communication has turned to human-machine communication. Also, the demand for an AI-based machine to recognize the user's emotions and give the corresponding feedback is becoming stronger. People expect the interactive machines to have the ability of observation, understanding, and abundant emotion similar to human beings, thus putting forward the new requirements for human-computer interaction [1]. However, the existing human-computer interaction mode of many service robots is mechanical and monotonous, relying only on the keywords matching and background search, which is not intelligent enough and lacks the understanding of the semantic context [2,3]. Therefore, we need to add emotional elements and intentional elements, and use affective computing technology to achieve emotional interaction. Emotional interaction has become the main trend in the human-computer interaction in the advanced information age. Besides, emotional interaction makes the human-computer interaction more intelligent. Namely, it makes the human-machine interaction as natural, cordial, vivid, emotional and temperature as human-human interaction is, thus realizing a deep human-computer interaction mode and understanding. Also, the emotion recognition plays an important role in the human-computer emotional interaction. Emotion recognition enables machines to perceive human emotional states and produce the ability of empathy. In the United States and Europe, many powerful laboratories have established special research groups to research and develop emotional systems and received sponsorship and support from some leading companies in that field. For instance, the famous Emotional Computing Team of the MIT Media Laboratory developed an emotional computing system, which collects data by using the biosensors and a camera capable of recording facial expressions; the collected data is then processed by the so-called "Emotional Assistant" adjustment program to recognize the human emotions [4]. Further, the Softbank company in Japan launched an emotional escort robot named Pepper, which can identify user emotions by analyzing facial expressions [5]. The Inner-scope, a neuroscience company, can predict whether the movie will make a splash by observing the highlights that make the audiences' brains highly active [6]. In [7], the authors propose a novel smart cushion system for detecting the user's stress state. In [8], the authors propose a novel emotional cognitive system, which can

* Corresponding authors.
E-mail addresses: yingyingjiang@hust.edu.cn (Y. Jiang), mshossain@ksu.edu.sa (M.S. Hossain), minchen@ieee.org (M. Chen).

https://doi.org/10.1016/j.inffus.2019.06.019
Received 18 February 2019; Received in revised form 6 June 2019; Accepted 9 June 2019
Available online 13 June 2019

analyze and predict postpartum depression based on prenatal data. In [9], the authors propose a creative gaming system to help users improvise. However, the current research on emotion recognition mostly focuses on single-modal recognition such as expression recognition, speech recognition, limb recognition, and physiological signal recognition. Nevertheless, the lack of the single-modal emotional information and vulnerability to various external factors lead to lower accuracy of emotion recognition (i.e., the facial expression is easily occluded, and speech is vulnerable to the interference from the surrounding-noise).

Emotion denotes a subjective attitude of humans nervous system toward the external relations. Brain first sends the instructions for the corresponding feedback which influences the human facial expression, frequency and speed of voice, and body language expressions, and also influences the human organs such as heart, arms, legs, brain, etc [10]. Therefore, considering a certain complementarity among different modal emotion data, researchers have started to use the facial expression, blink, gestures and some other psychophysical signals in the emotion recognition research. For example, in [11], the authors use three physiology signals, namely EDA, PPG and EMG, to identify human emotions together. Multimodal information fusion for data-driven emotion recognition has been attracting the attention of researchers in the affective computing filed. Compared to the single-mode emotion recognition, multimodal information fusion for data-driven emotion recognition has higher accuracy. The multimodal emotion data fusion and recognition was firstly proposed by Bigun and Duc in 1997 [12]. They fused the facial and voice data and put forward a statistical method based on the Bayesian theory. In recent years, the technology of artificial intelligence and muti-sensor data fusion [13,14] has been developing rapidly. Therefore, great progress has been achieved in research on the multimodal emotion data fusion and recognition. The multimodal emotion recognition has abundant and wide prospect of application. Besides, it helps to provide some useful functions to the empty-nest elderly and children [15]. By capturing human emotion, the psychological comfort for the empty-nest elderly and children can be achieved, helping to solve their psychological problems and undertake the load of psychologists. Through dialogue, a machine equipped with the mature artificial intelligence considers the patient's emotion and helps to alleviate disease.

With the aim to help those who are interested in the emotion recognition to know the multimodal emotion recognition comprehensively, we need to present a comprehensive and systematic survey. Although there existed a few review papers about multimodal emotion recognition, for example, the survey paper [1] analyzed the major trends and system-level factors correlated to the effects of multimodal emotion recognition. And paper [16] reviewed muti-sensor fusion. The review [17] mainly discussed the development history of affective computing, multimodal emotion dataset, methods for multimodal features extraction(such as visual, voice and textual features), multimodal fusion technology, and applicable API. However, the physiological signal (such as an EEG modality) was not considered. Due to the great improvement in the development of deep learning and some other AI technologies, many findings of the multimodal emotion recognition have been obtained in the last two years. The aforementioned papers don't cover the multimodal emotion recognition of AI technology fully. Also, we consider that the physiological change is take control of by the automatic nervous system and endocrine system, and almost not controlled by the subjective ideas, so the emotion recognition based on the physiological signal is objective [18]. Thus, with the change in the humans subtle physiological state (such as EEG and electrodermal activity), the specific fluctuations in human emotions can be observed and the corresponding emotional change can be recognized. For instance, when people become nervous under pressure or excited because of evil motive, the sympathetic nerve will cause the relevant somatic reactions, such as heartbeat acceleration, blood pressure increase, breath acceleration, body temperature rise, and even muscle or skin tremble [19–21]. Compared with the emotion recognition based on face recognition and movement, the recognition based on the EEG data or other physiological has higher credibility because it is natural and cannot be disguised or changed artificially. Besides, due to great achievements in the field of dry electrodes and wearable technology., the emotion analysis based on the EEG data obtained in a real environment (not limited to the laboratory environment) is more available [22,23].

Compared with the related literature [1,16,17], this paper mainly focuses on the data-driven multimodal emotion data fusion and recognition with AI technology. Considering the real-time emotion health monitoring system, the progress in key technologies related to the dataset, feature extraction, features fusion and classifying in the multimodal emotion recognition field is analyzed and summarized. This paper aims to comprehensively explain the data-driven multimodal emotion information fusion and help clearly understand the scientific problems and future research direction in that field.

2. Motivation example of multimodal information fusion

Fig. 1 shows the real-time emotion health surveillance system used in this paper. In this system, the following tasks are completed: the collection of multimodal emotion signal, labeling and selection of unlabeled emotion dataset on the edge cloud, multimodal emotion data fusion recognition and analysis of AI algorithm on the remote cloud, and the emotional feedback or decision-making control of the intelligent emotional interaction robot. A real-time personalized psychological health guardian can be offered to users.

2.1. Multimodal emotion data collection layer

The method for designing a high-efficiency emotion data collection is difficult in emotion recognition and interaction. The collected

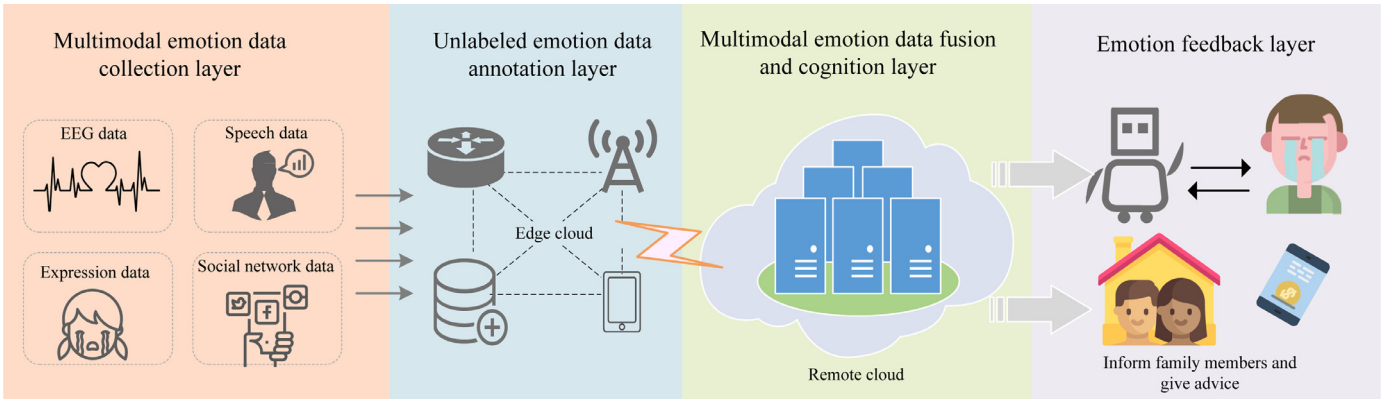

Fig. 1. An example of the real-time emotion healthcare system.

multimodal emotion data include the EEG data, voice data, expression data, and social contact information of users extracted from their smartphones. In the research findings of neurosciences and cognition science, emotional production has a high correlation with the physiological activity of the cerebral cortex. This offers the theoretical basis for recognizing user's emotion by researching the activities of his cerebral cortex (the EEG data) [24]. The collection of EEG data relies on wearable technology that has been developing rapidly. The 22-channel brain-wearable equipment designed independently is used in this work. An ADS1299-8 chip of TI company is used for EEG signal collection. And a CH559L is used as master chip. A comfortable, convenient, and non-intrusive brain-wearable equipment can be designed to collect user's EEG data in real time with high efficiency [25]. A MIC module and camera module are configured in an intelligent emotional interaction robot. They are respectively used to collect user's voice data and expression data on site in real time. User's voice, intonation, and expression can largely indicate user's psychological activity. As for the users who are good at concealing their emotion, the EEG data is very useful because it cannot be disguised or changed. Besides, thanks to the development of mobile network and intelligent terminals, a large number of users tend to announce their daily ideas and record their emotion on social networks [26]. The textual data of users they publish on social network reflects user's emotion change within a period. The above four types of emotion data are analyzed and modeled in this work to achieve higher emotion recognition accuracy.

2.2. Unlabeled emotion data labeling layer

The unlabeled learning algorithm deployed on the edge cloud can distinguish the validity of multimodal emotion data, upload important data and filter redundant data [27,28]. The remote cloud needs to recognize the user's emotion rapidly and accurately. Moreover, an intelligent robot with the ability of emotional interaction needs to make a decision as soon as possible and provide the corresponding emotional feedback to the user. The system collects user's multimodal emotion data in a large quantity. It is a great challenge to improve interaction delay and emotion cognition while maintaining the system's intelligence [29]. Therefore, the unlabeled learning algorithm introduced in [30] is used in this work. Instead of uploading a large number of original multimodal emotion data directly into the remote cloud, we will consider what effect the data will have on the data set after it is added to determine whether the unlabeled data is discarded or retained. Only the multimodal emotion data that can enhance the accuracy of emotion cognition will be uploaded into the remote cloud. In that way, the amount of uploaded data is decreased, and the intelligence of the remote cloud is maintained.

2.3. Multimodal emotion data fusion and cognition layer

After labeling and filtering multimodal emotion data on the edge cloud, data are uploaded to the remote cloud, and an AI algorithm is used to cognize and analyze the fusion of multimodal data [31]. The fusion of multimodal emotion data can help to get more comprehensive user's emotion features, enhance the robustness of emotion service system, and guarantee the system effective work when some emotion data lack. As for multimodal emotion data fusion, two key problems shall be solved: (1) how to effectively explore the relevance between different modalities and the emotion data that describe different modalities, and (2) how to fuse the emotion features or cognition results based on different modalities. Concretely, as for the EEG data, preprocessing removes the artifacts. The DBN (Deep Belief Network) algorithm is used for feature extraction. As for voice data, after the Mel spectrogram is obtained by preprocessing, the AlexNet DCNN (Deep Convolutional Neural Network) is used to extract the emotion features. As for facial expression data, the original facial expression images can be directly input into the VGGNet DCNN to extract the features of facial expression. As for textual data from the social network, the CNN (Convolutional Neural Network) is used to learn unstructured textual emotion data and extract the textual features. Finally, a feature fusion layer and a softmax classifier are designed in the neural network. The connection parameters are determined by supervised learning. Then the multimodal emotion data can be fused and cognized. Fig. 6 is the accuracy and loss function with different epoches.

2.4. Emotional feedback layer

Based on the fusion result of multimodal emotion data obtained from the remote cloud, the system can accurately cognize the user's emotions [32]. After cognizing the sorrowful and depressed emotion in a user, the intelligent robot can provide the corresponding emotion treatment through the emotional interaction and provide comfort to the user. For instance, the robot may play some music, say conciliative words to comfort the user, hug the user, and make the user feel empathy of the robot. Also, the system can send the information on a user's emotion to the mobile phones of user's family members and friends after the emotion is cognized. As for the depressed users, the symptom can be found as soon as possible, and medical suggestions should be given. The real-time emotional health monitoring system gives a personalized, intelligent, and humanized emotional feedback to respective users according to their characters and emotion status.

In the cases of real-time emotional health monitoring, the key problem is how to fuse the multimodal emotion data to increase the emotion cognition accuracy and offer an accurate real-time emotion service to users. As for the multimodal emotion cognition model deployed on the remote cloud, it is necessary to train and get a complete model using the existing dataset in advance. Besides, the model deployed on the remote cloud shall have good generalization ability on the multimodal data collected by several sensors in real time. A complete model trained with the open multimodal dataset has very important significance to the real-time emotional health monitoring system. Accordingly, the newest multimodal dataset, multimodal feature extraction, feature fusion and emotion classifier are discussed below.

3. Datasets

In most processes included in collecting the multimodal data for emotion cognition, the tested people are induced by videos or other means to generate certain emotions in people under the test. When wanted emotion is generated, the corresponding data is labeled and recorded. The recorded data is stored as a dataset called the induced or acted dataset. In some works, the spontaneous emotion of the tested users are recorded, these emotions were not stimulated by the external factor. Such a dataset is called the spontaneous dataset. Besides the old datasets which were surveyed in literature [17], such as HUMAINE [33], Belfast [34], SEMAINE [35], IEMOCAP [36] and eNTERFACE [37], many new achievements about multimodal datasets have been accomplished with the increase in user modalities. Therefore, this paper introduces the new five multimodal datasets presented in recent years. Table 1 shows the comparison of these six datasets.

AFEW: The dataset was collected by Abhinav et al. [38]. The authors collected the temporal videos from movies to depict real-world emotion expression as much as possible. The dataset mainly includes audio and visual modality. Two annotators to annotate the movie clips with a recommender system proposed in this paper. And the dataset have 330 subjects in total and seven main emotion label, namely sadness, happiness, disgust, anger, fear, surprise and the neutral class.

RECOLA: This dataset was put forward by Fabien et al. in 2013 [39]. The main modalities of this dataset are audio, visual, ECG, and EDA. It is spontaneous dataset based on the remote cooperative tasks. In addition, the emotion of participators is manipulated and balanced in dyads. There were 46 people included in the tests, of which 27 females and 19

Table 1
Comparison of multimodal emotion datasets.

Dataset	Modality	Subjects	Annotators	Emotion label
AFEW	Audio, visual	330	2	Happiness, sadness, anger, fear, disgust, surprise and neutral
RECOLA	Audio, visual, ECG, EDA	46	6	Arousal and valence
BAUM-1	Audio, visual	31	5 annotators for each clip	Happiness, anger, sadness, disgust, fear, surprise, boredom and contempt
EMOEEG	EEG, EOG, EMG, ECG, EDA	8	Self-assessment	Valence and arousal
CMU-MOSEI	Text, visual, audio	1000	3 crowdsourced judges	Happiness, sadness, anger, fear, disgust, surprise
WESAD	ECG, EDA, EMG, RESP, TEMP and ACC	15	Self-assessment	Neutral, stress, amusement

males, and the emotional labels are labeled by six annotaters according to two-dimensional coordinates.

BAUM-1: This dataset was put forward by Sara et al. [40] in 2017. It includes acted dataset and spontaneous dataset, denoted as BAUM-1a and BAUM-1s, repectively. The two main modalities are facial expression and voice. The emotion labels include happiness, anger, sadness, disgust, fear, surprise, boredom and contempt. The acted dataset is collected by asking the tested people to utter several sentences for the corresponding scene of eight imaged emotional labels. On the other hand, for the spontaneous dataset, a series of pictures and small videos stimulated the people under the test to generate target emotions. Two cameras were installed at the specific positions to capture the facial expression and voice of the tested people and record their emotion. There were in total 31 people under the test, including 13 females and 18 males at the age in the range 19–65. The total time for watching stimulating videos or pictures and conversing was 50 min for each participant. After getting the recorded videos, the videos were used to make clips by using the video processing technology. The acted dataset contained 273 clips while the spontaneous dataset contained 1184 clips. Each clip is annotated by five annotators, with scores ranging from 0 to 5. The higher the score is, the stronger the corresponding emotional state is. The voting method is used to vote on five points of each clip, and the final label is the one with the largest number of votes.

EMOEEG: This dataset was put forward by Anne-Claire et al. [41] at the 25th European Signal Processing Conference (EUSIPCO) in 2017. The main modality of the dataset is a physiological signal, including the EEG, EOG, EMG, ECG, EDA, etc. The Affectiva bracelet was used to record the skin conductance and temperature of the people during the test. There were in total eight people included in the test, of which 5 males and 3 females. The emotions were stimulated by the images and videos. Each image lasted for 25.5 s, and each video lasted for 28 s, and there were 25 image blocks, 50 videos, and 11 sessions. The emotional tagging was performed in a self-assessment way.

CMU-MOSEI: This dataset was built by Amir et al. [42] of CMU in 2018. It is the largest multimodal dataset at present, and it includes three modalities: text, video, and audio. The dataset contains 23,453 labeled videos from 1000 distinct speakers and covers 250 hot issues. Each video contains manual transcription that aligns audio and phoneme grade. The judges of the three Amazon Machinery Turkey platforms marked the video by crowdsourcing. Ekman criterion was used to classify emotions, namely happiness, sadness, anger, fear, disgust, and surprise.

WESAD: This dataset was put forward by Philip et al. [43] at the ICMI conference in 2018. It is a new open multimodal dataset aiming at the wearable pressure and emotion identification. This multimodal dataset contains the physiological data and sporting data collected by the equipment placed on participants' wrist and chest. There were 15 people included in the test, of which 12 males and 3 females at an average age of (27.5 ± 2.4). The data set consists of physiological data (ECG, EDA, EMG, RESP and TEMP) and exercise data (ACC). These high-resolution data were acquired by the device worn by the subjects' chest at a sampling rate of 700 Hz. Then subjects fill in self-report to complete emotional labeling. The self-reports represented the subjective experience during the emotional stimulus. The dataset contains three emotional statuses: neutral, stress, and amusement.

Table 2
EEG band range and decomposition level in [46].

Bandwidth	Frequency band	Decomposition level
1–4 Hz	δ	A6
4–8 Hz	θ	D6
8–10Hz	slow α	D5
8–12 Hz	α	
12–30 Hz	β	D4
30–40 Hz	γ	D3

4. Feature extraction

Different people have different emotion expression for the same emotion. Some people prefer to show their emotion with language, so their audio data contains more emotional clues [44]. On the other hand, some people tend to use facial expression to show their emotion. Also, some people are good at concealing and hiding their emotion, but their physiological features cannot be disguised or deceived. The main physiological features explored in this paper are the EEG features. Because of the great improvements in social networks, people tend to post their daily life and emotional status on the social network. Therefore, the multimodal emotion data fusion and cognition have begun to play an increasingly significant part for emotion recognition. Extraction of emotional features in different modalities is the key step of emotion cognition. The feature extraction technologies for different modalities have been deceived in recent years.

4.1. EEG features

Because the EEG signal is a very weak signal, it can be easily interfered by noise signals during the data collection process. Therefore, some preprocessing is needed to remove the artifacts from the EEG signals, including the electro-oculogram signal, electromyographic signal, etc. At present, the most frequently-used artifact removal methods are the filtering method and the independent component analysis [45].

According to the review in [46], the EEG emotion features are usually divided into frequency domain features, time domain features, and time-frequency features. Common time-domain features include event-related potential (ERP), statistical signals (such as average value, power, standard deviation, first-order deviation, normalized first-order deviation, and so on), non-stationary index (NSI), fractal dimension (DT), and higher order crossings (HOC). The frequency domain features mainly include band power, higher order spectra (HOS). The band characteristics of the EEG are generally decomposed into several frequency bands. Table 2 shows the range of commonly used frequency band, and decomposition level. Also, the frequency domain features are extracted from each frequency band, such as differential entropy (DE), power spectral density (PSD), etc. Lastly, common time-frequency features include Hilbert–Huang spectrum (HHS) and discrete wavelet transform (DWT).

Adrian et al. [47] combined the wavelet transform features, frequency domain features, and time domain features as the feature input for the EEG emotion cognition. The extracted features included the wavelet coefficients, maximum frequency amplitude, standard deviation, power, and mean value. Kairui Guo et al. [48] proposed to combine

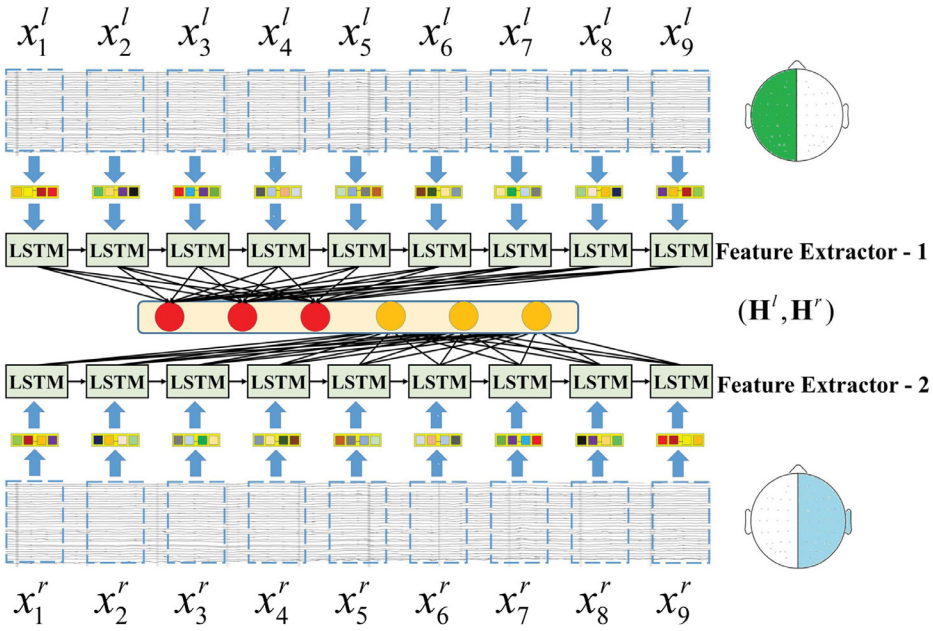

Fig. 2. The feature extractors of BiDANN in [52].

the time domain features and DWT features to build a new characteristic variable and then to combine the SVM and HMM as classifiers for the EEG emotion cognition. Concretely, first the relative wavelet energy was extracted, and then a relative wavelet entropy was extracted, and they were combined finally. The authors select standard deviation and DWT coefficients and multiply them to build a new characteristic variable.

In the recent latest years, due to the development in deep learning field, better performance has been achieved than by using the traditional machine learning methods for disposing of the problems of the EEG emotional features extraction and cognition. Zheng et al. [49] put forward the differential entropy features obtained from multiple EEG channels, and feed these features to DBN network, then train the network and extract more advanced emotional features. And because EEG data have high low-frequency energy in high-frequency energy, differential entropy has the balance ability to distinguish low-frequency EEG from high-frequency EEG. Specifically, the author uses a 1s long non-overlapping Hanning window and a short-time Fourier transform with a sampling point of 512 to extract the characteristics of each sub-band from the original EEG signal, and then obtain the differential entropy feature. In addition, the DBN and Hidden Markov Model were combined as auxiliary methods for getting more reliable emotion conversion status.

Mei et al. [50] and Samarth et al. [51] put forward to apply the CNN to extract the EEG emotional features and cognize emotions. This paper proves that the momentous information related to emotional state is contained in the functional connection matrix. The model proposed in this paper can extract the relevant features representing different emotions for learning.

Li et al. [52] proposed an original neural network model BiDANN used for extracting and cognizing the EEG emotional features, as shown in Fig. 2. The main content of BiDANN is mapping EEG signal corresponding to two hemispheres of the brain to discriminant the feature space respectively. Data could be classified easily. Distribution transformation between train set and test set as well as asymmetry of brain hemisphere were considered sufficiently in the model. The model contained two feature extractors which respectively learnt dynamic features of each brain hemisphere; the original EEG data was mapped to deep feature space with more discriminating emotion information. As for single brain hemisphere feature extractor, to use the time dependencein series sufficiently, a Long Short Term Memory (LSTM) framework was built to learn the contextual features and transfer the input to another space. Space was more effective and had components at a higher grade. By repeating the LSTM module, a series of hidden statuses completely representing input series were obtained; also, the feature space of the EEG was obtained.

Song et al. [53] presented a novel dynamic graph convolution neural network, which can dynamically learn the intrinsic relationship of various EEG channels through the adjacency matrix of the graph, thus discriminating the extraction of EEG emotional characteristics in different channels Specifically, the characteristics obtained from multiple EEG bands are used as input to the graph, and the channel of EEG is equivalent to the nodes in DGCNN graph. After the graph filtering, a 1×1 convolutional layer was employed for learning the features of discrimination among different frequency domains.

Lin et al. [54] proposed the Conditional Transfer Learning (cTL) for EEG emotion cognition. The model stimulated positive transfer of every individual (improved the subjects specificity without the increase of in the labeled data). The cTL first evaluates the transferability of the individual to the positive transfer, afterwards it optionally utilizes data from other people with comparable feature spaces. As for the original EEG signal of 30 channels and 30 s in every experiment, first the Short Fourier Transform was used with 50% overlapping 1-s Hamming window to transfer the signal to the frequency domain. The differential laterality was used to reflect the EEG spectral dynamics of emotional reaction in hemisphere frequency spectrum asymmetry signification. Then, the ReliefF method was exploited to obtain the features.

Choong et al. [55] put forward to use the Detrended Fluctuation Analysis (DFA) to detect the time relevance among the EEG signal features and complexity. When calculating DFA for feature extraction in each epoch, the minimum window size is 4 and the maximum window size is 76. The maximum window size is 1/10 of the epoch length, and the window size is incremented by four. Therefore, 19 windows of different size were analyzed in each stage.

Generally speaking, dimensionality reduction and selection of the EEG features are needed after the features are extracted. At present, Independent Component Analysis (ICA) [56], the Principal Component Analysis (PCA) [57], and Common Spatial Pattern (CSP) [58] are commonly used for that purpose.

4.2. Visual features

In the traditional machine learning methods, the facial expression features are extracted manually. In general, there are three group of methods for visual feature extraction: geometric methods based on the organs features and convex face positions, pixel methods based on textural face features, and mixing methods. Typical geometric methods include the Point Distribution Model (PDM) and Active Shape Model (ASM); pixel methods include the Gabor Wavelet, optical flow method, Scale invariant feature transformation (SIFT), local binary mode (LBP), linear discriminant analysis (LDA), etc; mixing methods include the Active Appearance Model (AAM), etc [59–61]. When a machine learning method is used to extract facial expression, it is not easy to manually design and build useful and effective features because it needs much professional field knowledge and effort. However, in recent years, with the rise of deep learning, many deep learning algorithms, such as convolutional neural network (CNN), have achieved good results in the field of visual recognition [62,63].

According to the survey presented in [64], deep learning tries to capture the high-grade features using several nonlinear transformations and layered architecture. In learning of emotional features of appearance, common deep learning models include CNN, DBN, Deep autoencoder (DAE), Recurrent neural network (RNN), and so on; among them, the CNN is used the most. In recent years, many researchers have used the CNNs to learn and recognize the emotional features of appearance.

In [65], the authors used a CNN architecture based on transfer learning. Firstly, the all-purpose pre-training of two CNN architectures with a different depth based on the ImageNet dataset was performed. They used two architectures, AlextNet and VGG-CNN-M-2018. In the first stage, the FER-2013 facial expression dataset was used for fine-tuning; then use the EmotiW dataset for the fine-tuning of the second phase, which allows the trained network weights to fit into the SFEW dataset. In the FaceNet2ExpNet proposed in [66], the authors designed a training algorithm with two stages and proposed a new distribution function for the high-level hidden neuron modeling of the expression network. Firstly, the paper pre-trains the convolutional layer and regularizes it with the mesh network to obtain the expression network. Then, the network obtained in the previous step is added to the fully connected layer network to extract the whole facial emotion feature. In [67], the authors not only demonstrated good CNN performance but also introduced a method to explain which face parts influence the CNN forecast. Visualization motivated the spatial pattern of different nerve cells at the convolutional layer in the largest degree to analyze network qualitatively and show the similarity to the facial action unit (FAU). In [68], several deep CNNs were trained as committee members, and their decisions were combined. There were two strategies in the model: (1) To get different decisions from the deep CNN, where the network structure and input normalization were changed at the beginning of training deep networks. (2)In order to obtain a better-performing committee in the aspects of structure and decision, a committee layer structure with exponential weighting and decision fusion was built. In [69], several CNNs were fused for learning the facial expression features. In [70], the proposed method contained the face detection module based on three existing technologies and the classification module with several deep CNNs, where each CNN model is trained on FER2013 data set first, and then fine-tuned on SFEW2.0 data set to further learn facial emotional features. In [71], The authors propose a deep neural network with two models for learning facial emotion features together. One of the networks extracts the appearance of the face from the original series of consecutive face images, and the other network extracts the temporal geometric features from the temporal face markers. Then joint fine-tuning was used for combination the two models. In [72], the authors propose a new action unit selection method to select the feature maps of the pre-trained CNN, analyzed all the feature maps from the 5th layer of the AlexNet and its relation with action unit (AU) and assessed their importance in facial expression recognition by feature ablation experiment. In [73] and [74], the CNN was used to extract the 3D facial expression features. In [75] and [76], several CNNs were aggregated to train the local characteristics and holistic characteristics of facial expression respectively. In [77], the authors combine RNN and CNN to propose a joint network architecture, which can extract temporal sequence features of dynamic facial expressions and static spatial features respectively, as shown in Fig. 3. In [78–81], the authors also combined CNN and RNN to learn the expression features, achieving a good effect.

In [82], the authors proposed a visual/video emotion cognition method through transfer learning. In [83], a new Boosted Deep Belief Net (BDBN) was proposed for executing three training stages iteratively in a uniform loop frame. The joint fine-tuning in the BDBN frame was used to improve the ability to judge features and strengthen their relative importance to the strong classifier. In [84], according to the sparsity of biological correlation and the cyclic characteristics of the network, an S-DSRN network is proposed to learn the emotional features of human

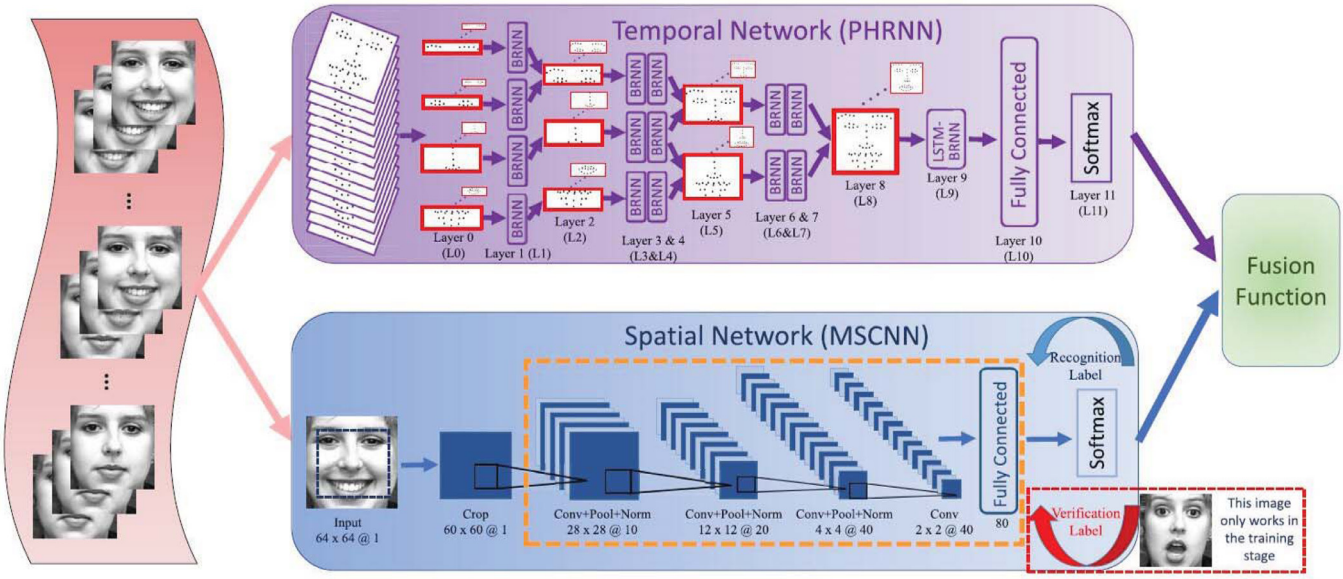

Fig. 3. The Spatial-Temporal Networks for facial expression feature extraction in [77].

faces, and the network can improve the stability of recognition results. The sparsity of features was obtained by using loss learning instead of additional penalty terms which are usually manually manufactured for sparse data representation in the proposed DSRN. In [85], The authors use a generative-contrastive network to learn facial emotional features. The network consists of three parts: generative network, contrastive network and discriminative network. The generated network generates reference pictures, and the contrastive network compares the real pictures with the generated pictures to obtain the contrastive features.

4.3. Audio features

At present, the acoustic features used for voice emotion recognition are divided into be prosodic features, relevant features based on a spectrum and voice quality features [86]. All the extraction of these features are usually done at the initial frame level. Concretely, the prosodic features include duration, pitch, and energy; the relevant features based on the spectrum typically include OSALPC (Unilateral Autocorrelation Linear Predictor Coefficient), MFCC (Mel Frequency Cepstral Coefficient), OSALPCC (OSALPC Based on Cepstral), LPC (Linear Predictor Coefficient), LPCC (Linear Predictor Cepstral Coefficient), LFPC (Logarithmic Frequency Power Coefficient), and so on [87]. The above three types of acoustic features can be fused to get the mixed features. They are low-level features extracted from those at the frame-level. For instance, in [88], the pitch, formants, zero crossing, MFCC, and its statistic parameters were mixed for extracting the acoustic features. In [89], the authors use cross-correlation model to calculate the correlation between the audio samples to be predicted and the audio samples with known labels, and then give the results of emotion recognition. Among them, the authors choose the six distinct mixed features of volume, zero crossing rate, energy, MFCC, spectral centroid and formant as emotional features for analysis.

Like the extraction method for emotion expression features, the application of deep learning for automatically extracting and learning voice emotion features has achieved a good effect. When deep learning is used for voice emotion recognition, the low-level features of each frame are usually extracted based on the traditional machine learning method, and next the high-level features can be extracted with deep learning automatically. In [90], the authors first divide the entire utterance into a series of segments and then get the primary audio features of each segment. The feature vectors corresponding to each spectrogram have MFCC, pitch, and delta features with temporal characteristics. Then, a DNN is used to build the utterance-level features from segment-level probability distributions. In [91], the authors pooled the last hidden layer and encoded each utterance to be a fixed-length vector. The process of feature coding is designed to use discourse level classifier for training in order to better classify. In [92], the emotional features of voice were learnt by combining a low-level feature extractor based on the Gaussian Mixed Model (GMM) with a high-level feature extractor based on DNN.

Many researchers used a CNN to extract the high-level features of voice. In [93–95], the authors first calculate low-level spectrums from each frame and then extracted the high-level features from an independent frame regarding spectrograms as the CNN input. In [96], the authors propose a new speech signal processing model that is inspired by the retina and convex lens imaging principle. According to the different distances between the spectrum map and the convex lens, the spectrum features of different size and different training data were obtained. Then, a CNN algorithm (AlextNet algorithm) was used to obtain the audio emotion features. In [97], the combination of phoneme and spectrogram features was used as the input of a multichannel CNN, and a good effect was achieved on the IEMOCAP dataset. In [98], a CNN was combined with an autoencoder to learn to distinguish the voice features which influenced emotion recognition. This model had two stages. In the first stage, a sparse automatic encoder (SAE) variable with the reconstitution punishment was used, and the unlabeled samples were used to learn the local invariant features (LIFs). Next, the LIFs were applied as an input of a feature extractor for the striking discerning feature analysis (SDFA). Then a new objective function is used to optimize the network and calculate the loss value. The emotional features learned from this function are significantly different, and they are orthogonal and discriminatory for speech emotion classification. However, the time series in speech are neglected in these methods of learning emotional features using CNN.

To solve the above mentioned time sequence problem of voice, many researchers used an RNN or an LSTM to extract and automatically learn the high-level features. In [99], the authors use the bidirectional long-term and short-term memory (BLSTM) model to extract temporal dynamic emotion features of speech. In [100], an RNN was employed to learn short frame acoustic features relevant to emotion and aggregate the features to be a compact utterance-level representation. Besides, a new feature pooling strategy was proposed. The focus was on the specific area of voice signal containing the emotion. In [101], the authors first extract 238 LLD features of the original speech signal, and then input them into the CTC-RNN network proposed in the paper. The network can automatically align the emotional segments of speech with the emotional tags, rather than the emotional segments with the non-emotional tags. In [102], the authors propose a novel pooling method based on the modulation spectrum, which can alleviate the influence of noisy background on emotional feature extraction. Then, the Multilayer Perceptron (MLP) and LSTM were used to get features of high level. In [103], the authors combined the CNN and LSTM. The phonetic tract length perturbation was first used for data enhancement, and the CNN was employed to get advanced audio emotional features from the spectrograms; while the Bi-LSTM was used to aggregate the long-term dependencies.

All the above methods for extracting the low-level features depend on the handcrafted features. To extract the features from the original phonetic oscillogram automatically, some researchers combined a CNN with an LSTM to extract the phonetic emotional features automatically [104,105]. In [106], the authors propose an end-to-end network architecture consisting of a CNN plus two layers of LSTM. The CNN is used to extract emotional features from the original speech signal, and then input the obtained features into the LSTM network to further learn the context features of the speech, thereby obtaining a complete speech emotion feature.

4.4. Text features

Textual information features extraction denotes the grammatical analysis and semantic analysis of text. By splitting the sentences, removing redundant information, settled words, participles and marked words, the emotional words that express the textual emotion tendency can be extracted. The traditional extraction of textual emotional features mainly depends on the rules-based technique, the Bag of Words (BoW) and some statistic methods [17]. In [107], the authors not only used the BoW to extract the textual emotional features but also proposed a new feature representation method named the eVector. However, after extracting the textual emotional features, the feature selection is needed. Feature selection refers to selecting the most suitable and effective features from the trained text features for further analysis. The frequency-used methods for feature selection are word frequency method (WF), document frequency method (DF), mutual information method (MI), information gain method, chi-square test method (CHI), etc [108,109].

Recently, scholars have tried to get feature representations from text data automatically. With the development of deep learning, the new deep learning methods have attracted more and more attention in the field of textual feature extraction [110]. In [111], the authors train a CNN model to extract the emotional features of the text. The input of the model is a vector of 306 dimensions per word. The convolution kernel of CNN extracts features by computing the semantically related word vectors and convolution layers in a hierarchical manner. In [112], the

authors use a CNN with the multiple resolution to identify the emotion of text information. The CNN was comprised of several parallel convolutions with the kernels of a different size to utilize textual information at different grades. In [113], the authors first extract the semantic word vector based on the word2vec method. At the same time, the authors map the emotional words in the text to the emotional space according to the affective lexicon to get the emotional word vector, and then gets the bottleneck feature of the emotional word vector based on autoencoder algorithm. Next the bottleneck features of semantic word vector and emotional word vector are fused to get the primary text features. Finally, the LSTM algorithm is used to learn more advanced text features.

5. Multimodal fusion and classification

The challenge of fusing the multimodal emotion data is brought by the emotion cognition and analysis after combining the heterogeneous emotion data modalities of different source and different time scale [114]. The fusion of multimodal emotion data can offer more reference information for emotion decisions, which raises the accuracy of general decisions. At present, there are mainly two kinds of multimodal fusion methods: feature-level fusion, and decision-level fusion. In some researches, model-level fusion is also introduced [115]. But compared with the former two fusion methods, model-level fusion is used relatively less. Therefore, this paper mainly introduced feature-level fusion and decision-level fusion with the development of AI.

5.1. Feature-level fusion

At present, feature level fusion is the most commonly used method in multi-modal emotion recognition, which connects features extracted from each modal into a new feature vector in some way. This new feature vector tends to have a higher dimension, and then uses the dimension reduction method, and finally uses a classifier to identify the emotion. The mutual relations between different modalities are used for feature layer fusion. However, the difference between the emotional features of different modalities is not considered. Also, time synchronism of different modalities is difficult to be achieved. With the increase of modalities, it is even more difficult to learn the relevance among modalities [17]. The traditional fusion methods include the concatenation, outer-product, etc.

In [116], data was input in different modalities to hidden units and time pooling units respectively, and then the features were fused in a multimodal fusion layer. Next, the input features were fused into the LSTM network for training. Finally, a linear regression layer was used for emotion recognition. In [117], the phonetic features in 1280 dimensions were extracted from the video while the features in 2048 dimensions were extracted from the images. Then, the two kinds of features were merged to obtain the feature vector in 3328 dimensions, which was fed to the 2-layer LSTM network for features training and recognition; a good effect was achieved on the RECOLA dataset. In [111], the phonetic, textual and image features were directly connected for feature fusion after the features had been extracted and selected. Then, the MKL classifier was used for emotion recognition. In [118], a new method on the basis of integration for visual and phonetic features following the bilinear pooling theory was proposed; the DBN was used to train the fused features; finally, a softmax classifier was employed to recognize emotions. In [119], a new feature fusion method based on the Relational Tensor Network which fused text, audio, and image, was proposed. The validity of this method was verified on the CMU-MOSEI dataset. In [120], the authors use self-attention mechanism to fuse audio features and text features to get new emotional features. In [121] and [122], the authors use the DBN network to fuse the speech emotional features and the facial emotional features, and then trains them to learn the new emotional features after the fusion of the two modals; finally, the SVM was used for emotion classification and recognition, as shown Fig. 4. In [123], the authors use extreme learning machines (ELM) to fuse the features of speech and video in feature level. Specifically, the Mel spectrogram of the speech is first input to the 2D CNN and the key frames of the video are input into the 3D CNN to extract the features of the fully connected layer, and then the features are merged by two consecutive ELMs. The first ELM contains 100 hidden units and the second ELM contains 250 hidden units. Then the output of the second ELM is continuously put into softmax and SVM for emotional recognition to get the final result.

5.2. Decision-level fusion

In the general process of decision layer fusion, a respective classifier considering the EEG features, phonetic features, textual features, and expression features is used for emotion recognition. Then, an algebraic combination rule is adopted to combine single-modality emotion recognition results to improve the recognition accuracy of final result. Compared with the feature layer fusion, the decision layer fusion emphasizes the difference between different features. The most suitable classifier can be chosen for each modality. However, the relevance between features is not considered. Also, the learning process of learning is long and time-consuming [17].

In [124], the weighted product rule was adopted to fuse the audio and image recognition results at a decision layer. Specifically, SVM is used to classify each feature, and then the weights in the fusion network are multiplied by the probabilistic values for each class of each feature obtained before, and then the values belonging to the same class are added together. Eventually, the final label is selected which has the greatest probabilistic value. In [125], the EEG and facial expression were respectively classified, and the sum rule and production rule were used to fuse the recognition results. In [126], a quality adaptive multimodal fusion scheme was used for multimodal data integration of the ECG, EEG, CSR and facial expression at the decision layer. Then the final output result is obtained by fusing the results of each mode according to the fusion method of decision level, as shown in Fig. 5. Moreover, the authors released their dataset, namely QAMAF online. In [127], the authors use decision-level fusion to integrate various network models. The paper first extracts audio features and continuous image sequences from video. After the basic preprocessing operation of images, the recognition results are obtained by using CNN-RNN and C3D network respectively. At the same time, SVM is used to recognize the preprocessed audio data. The predicted scores obtained from the different models are blended by weighted summation.

5.3. Findings and discussion

Studies about feature-level fusion and decision-level fusion in this paper are summarized in Table 3. We find that in the current multimodal emotion data fusion technology, feature-level fusion is used more than decision-level fusion. The feature extraction for each modality is mostly based on AI algorithms such as deep learning. Then, for the high-dimensional features after fusion, the pooling method or other dimensionality reduction methods such as PCA are exploited to select features and reduce the feature dimensions. Then the deep learning method is used for training. Finally, a softMax classifier or a simple linear classifier such as SVM is used for emotion classification. In decision-level fusion strategy, sum-rule and product-rule are major trend. In addition, for emotion label, discrete basic emotions (happiness, anger, sadness, disgust, fear, and surprise) are used more than dimension methods (arousal and valence).

In Section 2, we proposed a motivation example of multimodal information fusion for data-driven emotion recognition. According to the feature extraction method of each modal mentioned above and the multimodal fusion strategy, we have conducted experiment for the motivation example. In this experiment, we use DBN algorithm to extract EEG features, AlexNet DCNN network to extract advanced features from Mel spectrum of speech, VGGNet DCNN network to extract features from

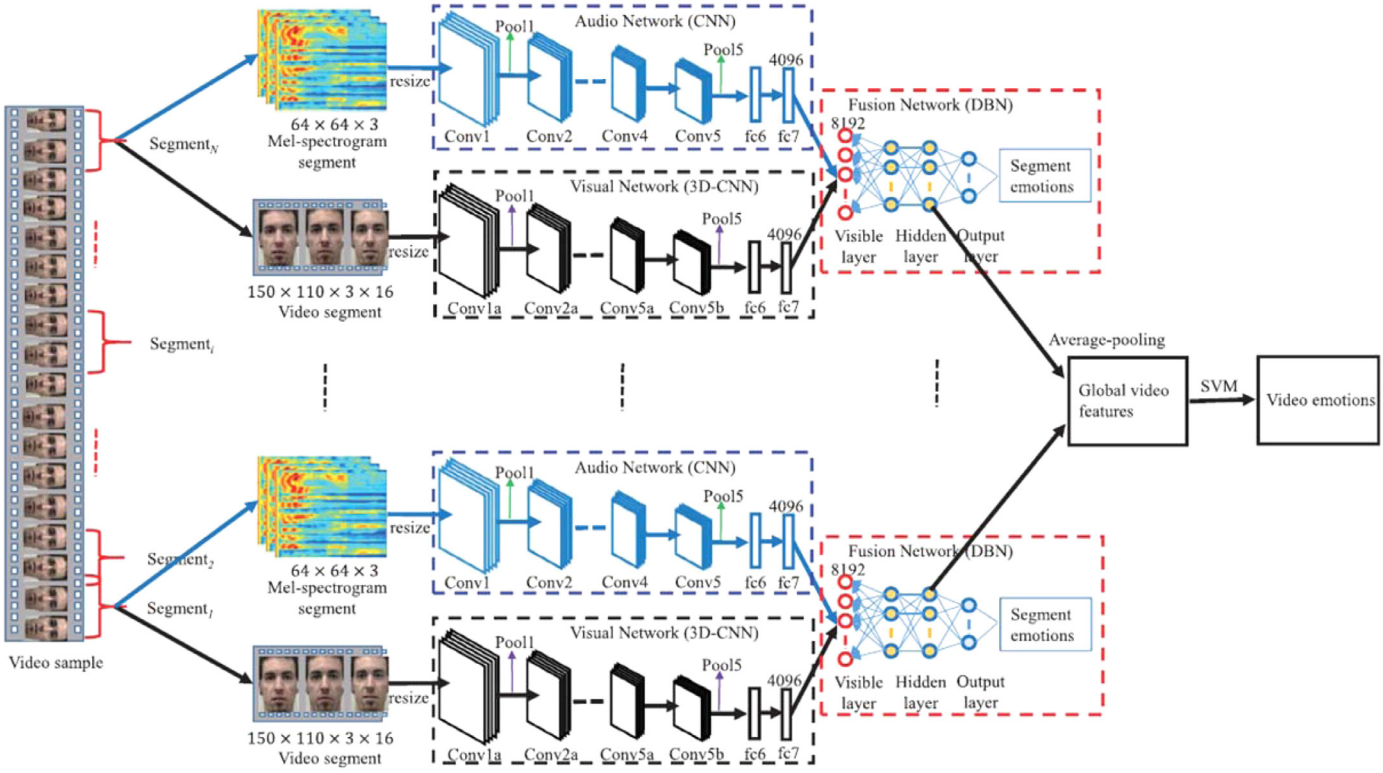

Fig. 4. Feature layer fusion in [121].

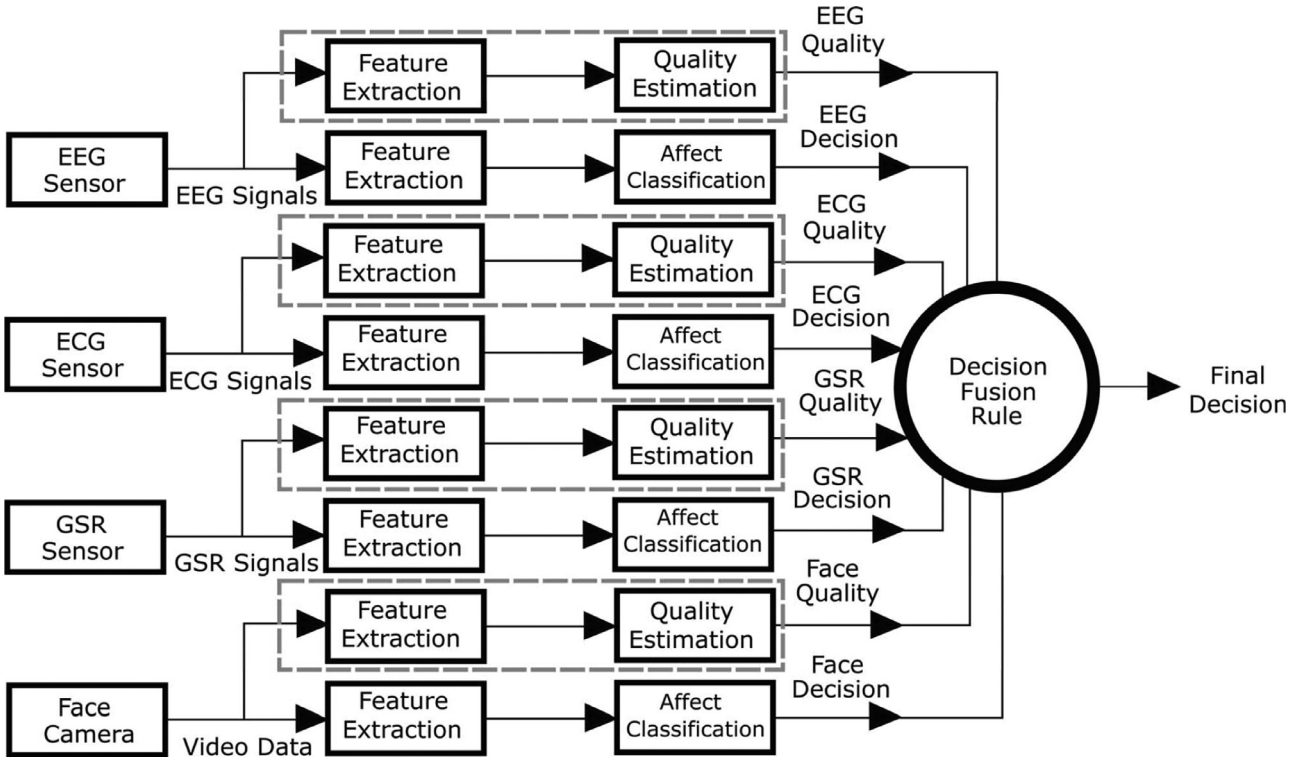

Fig. 5. Decision layer fusion presented in [126].

Table 3
Comparison of accuracy with different fusion methods.

Fusion method	Dataset	Reference	Modality	Average accuracy
Feature-level	RECOLA	Chao et al. [116]	Audio, visual, ECG, EDA	0.667
		Tzirakis et al. [117]	Audio, visual	0.760
	IEMOCAP	Poria et al. [111]	Audio. visual, text	0.776
		Hazarika et al. [120]	Audio, text	0.721
	eNterface	Nguyen et al. [118]	Audio, visual	0.9085
		Zhang et al. [121]	Audio, visual	0.8597
	CMU-MOSEI	Sahay et al. [119]	Audio, visual, text	0.4917
	BAUM-1s	Zhang et al. [121]	Audio, visual	0.5457
Decision-level	AFEW	Sun et al. [124]	Audio, visual	0.512
		Fan et al. [127]	Audio, visual	0.5902
	QAMAF	Gupta et al. [126]	EEG,ECG,GSR, visual	0.59

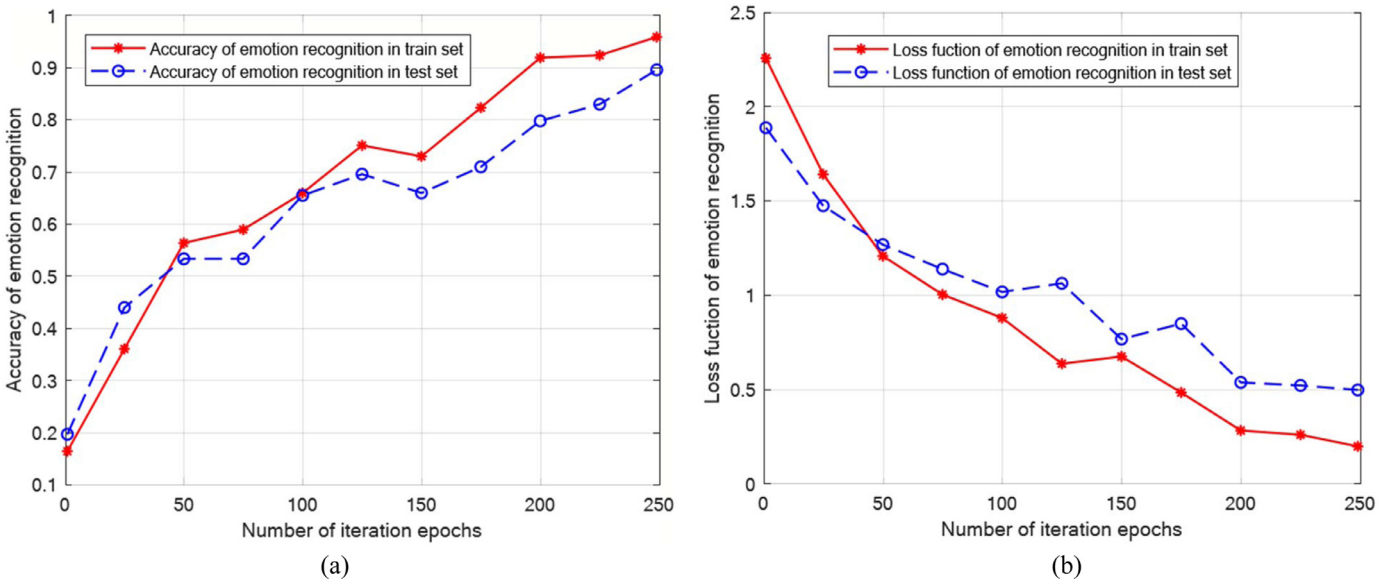

Fig. 6. The experiment result of emotion recognition: (a) Accuracy of emotion recognition with the number of iteration epochs; (b) Loss function of emotion recognition with the number of iteration epochs.

user's facial expressions, and CNN to extract features from text content of social network. Then the feature-level fusion strategy proposed in [121] is used to fuse the features of each modal. Finally, the softmax classifier is used to classify the emotions. Fig. 6 is the accuracy and loss function with different epoches. We can see that with the increase of iterations, the accuracy of training set and validation set increases, the value of loss function decreases, and the accuracy finally converges to about 90%, which achieves better results.

6. Conclusion

Taking the real-time emotional health monitoring systems as an example, this paper comprehensively reviews and summarizes the relevant key technologies in the field of multimodal information fusion for data-driven emotion recognition. The existing extraction technologies for the open dataset, EEG, audio, visual and textual features, feature layer fusion, decision layer fusion and classification are discussed in details. These presented discussions aim to provide a comprehensive overview and a big-picture of this exciting and hot-spot research area.

Acknowledgment

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University , Riyadh, Saudi Arabia for funding this work through the research group project no. RGP -318.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2019.06.019.

References

[1] S.K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Comput. Surv. (CSUR) 47 (3) (2015) 43.
[2] M. Chen, P. Zhou, D. Wu, L. Hu, M. Mehedi Hassan, H. Atif Alamri, Ai-skin : skin disease recognition based on self-learning and wide data collection through a closed loop framework, arXiv:1906.01895 (2019).
[3] Y. Qian, Y. Zhang, X. Ma, H. Yu, L. Peng, Ears: emotion-aware recommender system based on hybrid information fusion, Inf. Fusion 46 (2019) 141–146.
[4] C.L. Breazeal, Designing Sociable Robots, MIT press, 2004.
[5] S. Robotics, Pepper, Softbank Robot. (2016).
[6] T. Bartelme, Meet carl marci: a doctor who wants to measure your emotions, Physician Exec. 38 (1) (2012) 10.
[7] R. Gravina, Q. Li, Emotion-relevant activity recognition based on smart cushion using multi-sensor fusion, Inf. Fusion 48 (2019) 1–10.
[8] M.W. Moreira, J.J. Rodrigues, N. Kumar, K. Saleem, I.V. Illin, Postpartum depression prediction through pregnancy data analysis for emotion-aware smart systems, Inf. Fusion 47 (2019) 23–31.
[9] M. Chen, Y. Jiang, Y. Cao, Y.A. Zomaya, Creativebioman: brain and body wearable computing based creative gaming system, arXiv:1906.01801 (2019).
[10] C.E. Izard, Human Emotions, Springer Science & Business Media, 2013.
[11] 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.
[12] B. Duc, E.S. Bigün, J. Bigün, G. Maître, S. Fischer, Fusion of audio and video information for multi modal person authentication, Pattern Recognit. Lett. 18 (9) (1997) 835–843.

[13] Z. Wang, D. Wu, R. Gravina, G. Fortino, Y. Jiang, K. Tang, Kernel fusion based extreme learning machine for cross-location activity recognition, Inf. Fusion 37 (2017) 1–9.
[14] G. Fortino, S. Galzarano, R. Gravina, W. Li, A framework for collaborative computing and multi-sensor data fusion in body sensor networks, Inf. Fusion 22 (2015) 50–70.
[15] M. Chen, Y. Ma, J. Song, C.-F. Lai, B. Hu, Smart clothing: connecting human with clouds and big data for sustainable health monitoring, Mobile Networks Appl. 21 (5) (2016) 825–845.
[16] R. Gravina, P. Alinia, H. Ghasemzadeh, G. Fortino, Multi-sensor fusion in body sensor networks: state-of-the-art and research challenges, Inf. Fusion 35 (2017) 68–80.
[17] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: from unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.
[18] 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.
[19] M. Khezri, M. Firoozabadi, A.R. Sharafat, Reliable emotion recognition system based on dynamic adaptive fusion of forehead biopotentials and physiological signals, Comput. Methods Programs Biomed. 122 (2) (2015) 149–164.
[20] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, M. Pantic, Av+ ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data, in: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, ACM, 2015, pp. 3–8.
[21] M.K. Abadi, R. Subramanian, S.M. Kia, P. Avesani, I. Patras, N. Sebe, Decaf: meg-based multimodal database for decoding affective physiological responses, IEEE Trans. Affect. Comput. 6 (3) (2015) 209–222.
[22] 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.
[23] H. Ghasemzadeh, P. Panuccio, S. Trovato, G. Fortino, R. Jafari, Power-aware activity monitoring using distributed wearable sensors, IEEE Trans. Hum. Mach. Syst. 44 (4) (2014) 537–544.
[24] T. Dalgleish, The emotional brain, Nat. Rev. Neurosci. 5 (7) (2004) 583.
[25] M. Chen, J. Zhou, G. Tao, J. Yang, L. Hu, Wearable affective robot, IEEE Access 6 (2018) 64766–64776.
[26] G. Muhammad, M.F. Alhamid, User emotion recognition from a larger pool of social network data using active learning, Multimed. Tools Appl. 76 (8) (2017) 10881–10892.
[27] M. Chen, W. Li, G. Fortino, Y. Hao, L. Hu, I. Humar, A dynamic service migration mechanism in edge cognitive computing, ACM Trans. Internet Technol. (TOIT) 19 (2) (2019) 30.
[28] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, R. Jafari, Enabling effective programming and flexible management of efficient body sensor network applications, IEEE Trans. Hum. Mach. Syst. 43 (1) (2013) 115–133.
[29] X. Chen, Y. Zhao, Y. Li, QoE-aware wireless video communications for emotion-aware intelligent systems: a multi-layered collaboration approach, Inf. Fusion 47 (2019) 1–9.
[30] M. Chen, Y. Hao, K. Lin, Z. Yuan, L. Hu, Label-less learning for traffic control in an edge network, IEEE Netw. 32 (6) (2018) 8–14.
[31] G. Smart, N. Deligiannis, R. Surace, V. Loscri, G. Fortino, Y. Andreopoulos, Decentralized time-synchronized channel swapping for ad hoc wireless networks, IEEE Trans. Veh. Technol. 65 (10) (2016) 8538–8553.
[32] G. Fortino, D. Parisi, V. Pirrone, G. Di Fatta, Bodycloud: a saas approach for community body sensor networks, Future Gen. Comput. Syst. 35 (2014) 62–79.
[33] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, et al., The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data, in: International conference on affective computing and intelligent interaction, Springer, 2007, pp. 488–500.
[34] E. Douglas-Cowie, R. Cowie, M. Schröder, A new emotion database: considerations, sources and scope, ISCA tutorial and research workshop (ITRW) on speech and emotion, 2000.
[35] 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) (2012) 5–17.
[36] 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.
[37] 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.
[38] A. Dhall, R. Goecke, S. Lucey, T. Gedeon, et al., Collecting large, richly annotated facial-expression databases from movies, IEEE Multimedia 19 (3) (2012) 34–41.
[39] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the recola multimodal corpus of remote collaborative and affective interactions, in: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on, IEEE, 2013, pp. 1–8.
[40] 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) (2017) 300–313.
[41] A.-C. Conneau, A. Hajlaoui, M. Chetouani, S. Essid, Emoeeg: a new multimodal dataset for dynamic eeg-based emotion recognition with audiovisual elicitation, in: Signal Processing Conference (EUSIPCO), 2017 25th European, IEEE, 2017, pp. 738–742.
[42] A.B. Zadeh, P.P. Liang, S. Poria, E. Cambria, L.-P. Morency, Multimodal language analysis in the wild: cmu-mosei dataset and interpretable dynamic fusion graph, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1, 2018, pp. 2236–2246.
[43] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, K. Van Laerhoven, Introducing wesad, a multimodal dataset for wearable stress and affect detection, in: Proceedings of the 2018 on International Conference on Multimodal Interaction, ACM, 2018, pp. 400–408.
[44] M. Chen, F. Herrera, K. Hwang, Cognitive computing: architecture, technologies and intelligent applications, IEEE Access 6 (2018) 19774–19783.
[45] J. Preethi, M. Sreeshakthy, A. Dhilipan, A survey on eeg based emotion analysis using various feature extraction techniques, Int. J. Sci. Eng.Technol. Res. (IJSETR) 3 (11) (2014).
[46] R. Jenke, A. Peer, M. Buss, Feature extraction and selection for emotion recognition from eeg, IEEE Trans. Affect. Comput. 5 (3) (2014) 327–339.
[47] A.Q.-X. Ang, Y.Q. Yeong, W. Wee, Emotion classification from eeg signals using time-frequency-dwt features and ann, J. Comput. Commun. 5 (03) (2017) 75.
[48] K. Guo, H. Candra, H. Yu, H. Li, H.T. Nguyen, S.W. Su, Eeg-based emotion classification using innovative features and combined svm and hmm classifier, in: Engineering in Medicine and Biology Society (EMBC), 2017 39th Annual International Conference of the IEEE, IEEE, 2017, pp. 489–492.
[49] W.-L. Zheng, J.-Y. Zhu, Y. Peng, B.-L. Lu, Eeg-based emotion classification using deep belief networks, in: Multimedia and Expo (ICME), 2014 IEEE International Conference on, IEEE, 2014, pp. 1–6.
[50] H. Mei, X. Xu, Eeg-based emotion classification using convolutional neural network, in: Security, Pattern Analysis, and Cybernetics (SPAC), 2017 International Conference on, IEEE, 2017, pp. 130–135.
[51] S. Tripathi, S. Acharya, R.D. Sharma, S. Mittal, S. Bhattacharya, Using deep and convolutional neural networks for accurate emotion classification on deap dataset, in: AAAI, 2017, pp. 4746–4752.
[52] Y. Li, W. Zheng, Z. Cui, T. Zhang, Y. Zong, A novel neural network model based on cerebral hemispheric asymmetry for eeg emotion recognition, in: IJCAI, 2018, pp. 1561–1567.
[53] T. Song, W. Zheng, P. Song, Z. Cui, Eeg emotion recognition using dynamical graph convolutional neural networks, IEEE Trans. Affect. Comput. (2018).
[54] Y.-P. Lin, T.-P. Jung, Improving eeg-based emotion classification using conditional transfer learning, Front. Hum. Neurosci. 11 (2017) 334.
[55] W. Choong, W. Khairunizam, M. Omar, M. Murugappan, A. Abdullah, H. Ali, S. Bong, Eeg-based emotion assessment using detrended flunctuation analysis (dfa), J. Telecommun. Electron.Comput. Eng. (JTEC) 10 (1–13) (2018) 105–109.
[56] A. Hyvärinen, J. Karhunen, E. Oja, Independent component analysis, 46, John Wiley & Sons, 2004.
[57] I. Jolliffe, Principal component analysis, Springer, 2011.
[58] K.K. Ang, Z.Y. Chin, H. Zhang, C. Guan, Filter bank common spatial pattern (fbcsp) in brain-computer interface, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), IEEE, 2008, pp. 2390–2397.
[59] B. Fasel, J. Luettin, Automatic facial expression analysis: a survey, Pattern Recognit. 36 (1) (2003) 259–275.
[60] G. Sandbach, S. Zafeiriou, M. Pantic, L. Yin, Static and dynamic 3d facial expression recognition: a comprehensive survey, Image Vis. Comput. 30 (10) (2012) 683–697.
[61] 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.
[62] M. Moghimi, S.J. Belongie, M.J. Saberian, J. Yang, N. Vasconcelos, L.-J. Li, Boosted convolutional neural networks, BMVC, 2016.
[63] M. Chen, X. Shi, Y. Zhang, D. Wu, M. Guizani, Deep features learning for medical image analysis with convolutional autoencoder neural network, IEEE Trans. Big Data (2017).
[64] S. Li, W. Deng, Deep facial expression recognition: a survey, arXiv:1804.08348 (2018).
[65] 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, ACM, 2015, pp. 443–449.
[66] H. Ding, S.K. Zhou, R. Chellappa, Facenet2expnet: regularizing a deep face recognition net for expression recognition, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 118–126.
[67] P. Khorrami, T. Paine, T. Huang, Do deep neural networks learn facial action units when doing expression recognition? in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015, pp. 19–27.
[68] B.-K. Kim, J. Roh, S.-Y. Dong, S.-Y. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, J. Multimodal User Interfaces 10 (2) (2016) 173–189.
[69] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, E. Xun, Ensemble of deep neural networks with probability-based fusion for facial expression recognition, Cognit. Comput. 9 (5) (2017) 597–610.
[70] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 435–442.
[71] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2983–2991.
[72] Y. Zhou, B.E. Shi, Action unit selective feature maps in deep networks for facial expression recognition, in: Neural Networks (IJCNN), 2017 International Joint Conference on, IEEE, 2017, pp. 2031–2038.

[73] H. Li, J. Sun, Z. Xu, L. Chen, Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network, IEEE Trans. Multimedia 19 (12) (2017) 2816–2831.
[74] A. Jan, H. Ding, H. Meng, L. Chen, H. Li, Accurate facial parts localization and deep learning for 3d facial expression recognition, in: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, IEEE, 2018, pp. 466–472.
[75] S. Xie, H. Hu, Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks, IEEE Trans. Multimedia 21 (1) (2019) 211–220.
[76] Y. Fan, J.C. Lam, V.O. Li, Multi-region ensemble convolutional neural network for facial expression recognition, in: International Conference on Artificial Neural Networks, Springer, 2018, pp. 84–94.
[77] K. Zhang, Y. Huang, Y. Du, L. Wang, Facial expression recognition based on deep evolutional spatial-temporal networks, IEEE Trans. Image Process. 26 (9) (2017) 4193–4203.
[78] P. Rodriguez, G. Cucurull, J. Gonzalez, J.M. Gonfaus, K. Nasrollahi, T.B. Moeslund, F.X. Roca, Deep pain: exploiting long short-term memory networks for facial expression classification, IEEE Trans. Cybern. (99) (2017) 1–11.
[79] B. Hasani, M.H. Mahoor, Facial expression recognition using enhanced deep 3d convolutional neural networks, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on, IEEE, 2017, pp. 2278–2288.
[80] N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, M. Zareapoor, Hybrid deep neural networks for face emotion recognition, Pattern Recognit. Lett. (2018).
[81] D.K. Jain, Z. Zhang, K. Huang, Multi angle optimal pattern-based deep learning for automatic facial expression recognition, Pattern Recognit. Lett. (2017).
[82] Y. Hao, J. Yang, M. Chen, M.S. Hossain, M.F. Alhamid, Emotion-aware video qoe assessment via transfer learning, IEEE Multimedia (2018).
[83] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1805–1812.
[84] M. Alam, L.S. Vidyaratne, K.M. Iftekharuddin, Sparse simultaneous recurrent deep learning for robust facial expression recognition, IEEE Trans. Neural Netw. Learn. Syst. (2018).
[85] Y. Kim, B. Yoo, Y. Kwak, C. Choi, J. Kim, Deep generative-contrastive networks for facial expression recognition, 2017 arXiv:1703.07140.
[86] W. Han, H. Li, H. Ruan, L. Ma, Review on speech emotion recognition, J.Software 25 (1) (2014) 37–50.
[87] W. Dai, D. Han, Y. Dai, D. Xu, Emotion recognition and affective computing on vocal social media, Inf. Manag. 52 (7) (2015) 777–788.
[88] P.P. Dahake, K. Shaw, P. Malathi, Speaker dependent speech emotion recognition using mfcc and support vector machine, in: Automatic Control and Dynamic Optimization Techniques (ICACDOT), International Conference on, IEEE, 2016, pp. 1080–1084.
[89] J. Chatterjee, V. Mukesh, H.-H. Hsu, G. Vyas, Z. Liu, Speech emotion recognition using cross-correlation and acoustic features, in: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), IEEE, 2018, pp. 243–249.
[90] K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, Fifteenth annual conference of the international speech communication association, 2014.
[91] Z.-Q. Wang, I. Tashev, Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks, in: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, 2017, pp. 5150–5154.
[92] I.J. Tashev, Z.-Q. Wang, K. Godin, Speech emotion recognition based on gaussian mixture models and deep neural networks, in: Information Theory and Applications Workshop (ITA), 2017, IEEE, 2017, pp. 1–4.
[93] S. Parthasarathy, I. Tashev, Convolutional neural network techniques for speech emotion recognition, in: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC), IEEE, 2018, pp. 121–125.
[94] A.M. Badshah, J. Ahmad, N. Rahim, S.W. Baik, Speech emotion recognition from spectrograms with deep convolutional neural network, in: Platform Technology and Service (PlatCon), 2017 International Conference on, IEEE, 2017, pp. 1–5.
[95] L. Zheng, Q. Li, H. Ban, S. Liu, Speech emotion recognition based on convolution neural network combined with random forest, in: 2018 Chinese Control And Decision Conference (CCDC), IEEE, 2018, pp. 4143–4147.
[96] Y. Niu, D. Zou, Y. Niu, Z. He, H. Tan, Improvement on speech emotion recognition based on deep convolutional neural networks, in: Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, ACM, 2018, pp. 13–18.
[97] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, J. Vepa, Speech emotion recognition using spectrogram & phoneme embedding, in: Proc. Interspeech 2018, 2018, pp. 3688–3692.
[98] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching, IEEE Trans. Multimedia 20 (6) (2018) 1576–1590.
[99] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition (2015).
[100] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, 2017, pp. 2227–2231.
[101] X. Chen, W. Han, H. Ruan, J. Liu, H. Li, D. Jiang, Sequence-to-sequence modelling for categorical speech emotion recognition using recurrent neural network, in: 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), IEEE, 2018, pp. 1–6.
[102] A.R. Avila, J. Monteiro, D. O'Shaughneussy, T.H. Falk, Speech emotion recognition on mobile devices based on modulation spectral feature pooling and deep neural networks, in: 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE, 2017, pp. 360–365.
[103] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, B. Schmauch, Cnn+ lstm architecture for speech emotion recognition with data augmentation, in: Proc. Workshop on Speech, Music and Mind 2018, 2018, pp. 21–25.
[104] 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: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, IEEE, 2016, pp. 5200–5204.
[105] W. Lim, D. Jang, T. Lee, Speech emotion recognition using convolutional and recurrent neural networks, in: Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific, IEEE, 2016, pp. 1–4.
[106] P. Tzirakis, J. Zhang, B.W. Schuller, End-to-end speech emotion recognition using deep neural networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 5089–5093.
[107] Q. Jin, C. Li, S. Chen, H. Wu, Speech emotion recognition with acoustic and lexical features, in: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 4749–4753.
[108] S.N. Shivhare, S. Khethawat, Emotion detection from text, arXiv:1205.4944 (2012).
[109] D. Ghazi, D. Inkpen, S. Szpakowicz, Hierarchical approach to emotion recognition and classification in texts, in: Canadian Conference on Artificial Intelligence, Springer, 2010, pp. 40–50.
[110] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, H. Prendinger, Deep learning for affective computing: text-based emotion recognition in decision support, Decis. Support Syst. 115 (2018) 24–35.
[111] S. Poria, I. Chaturvedi, E. Cambria, A. Hussain, Convolutional mkl based multimodal emotion recognition and sentiment analysis, in: Data Mining (ICDM), 2016 IEEE 16th International Conference on, IEEE, 2016, pp. 439–448.
[112] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, N. Dehak, Deep neural networks for emotion recognition combining audio and transcripts, in: Proc. Interspeech 2018, 2018, pp. 247–251.
[113] 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.
[114] 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.
[115] V. Vielzeuf, S. Pateux, F. Jurie, Temporal multimodal fusion for video emotion classification in the wild, in: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ACM, 2017, pp. 569–576.
[116] L. Chao, J. Tao, M. Yang, Y. Li, Z. Wen, Long short term memory recurrent neural network based multimodal dimensional emotion recognition, in: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, ACM, 2015, pp. 65–72.
[117] 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. Signal Process. 11 (8) (2017) 1301–1309.
[118] D. Nguyen, K. Nguyen, S. Sridharan, D. Dean, C. Fookes, Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition, Comput. Vision Image Understanding 174 (2018) 33–42.
[119] S. Sahay, S.H. Kumar, R. Xia, J. Huang, L. Nachman, Multimodal relational tensor network for sentiment and emotion classification, arXiv:1806.02923 (2018).
[120] D. Hazarika, S. Gorantla, S. Poria, R. Zimmermann, Self-attentive feature-level fusion for multimodal emotion detection, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2018, pp. 196–201.
[121] 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) (2018) 3030–3043.
[122] 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.
[123] M.S. Hossain, G. Muhammad, Emotion recognition using deep learning approach from audio–visual emotional big data, Inf. Fusion 49 (2019) 69–78.
[124] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, Q. Wei, Combining multimodal features within a fusion network for emotion recognition in the wild, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 497–502.
[125] Y. Huang, J. Yang, P. Liao, J. Pan, Fusion of facial expressions and eeg for multimodal emotion recognition, Comput. Intell. Neurosci. 2017 (2017).
[126] 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, ACM, 2016, pp. 317–320.
[127] 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, ACM, 2016, pp. 445–450.

Yingying Jiang (e-mail: yingyingjiang@hust.edu.cn) received her bachelor degree from School of Information and Safety Engineering, Zhongnan University of Economics and Law (ZUEL) in June 2017. Currently, she is a Ph.D candidate at Embedded and Pervasive Computing (EPIC) Lab in School of Computer Science and Technology, HUST. Her research interests include healthcare big data, congnitive learning, etc.

Wei Li is with the School of Computer Science and Technology, Huazhong University of Science and Technology, China (e-mail: weili_epic@hust.edu.cn). Wei Li is a Ph.D candidate at the School of Computer Science and Technology, Huazhong University of Science and Technology, China. Her research interests include software defined IoT, deep learning, etc.

M. Shamim Hossain (mshossain@ksu.edu.sa) is a Professor at the Department of Software Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. He is also an adjunct professor at the School of Electrical Engineering and Computer Science, University of Ottawa, Canada. His research interests include cloud networking, smart environment (smart city, smart health), social media, IoT, edge computing and multimedia for health care, deep learning approach to multimedia processing, and multimedia big data. He has authored and coauthored more than 200 publications including refereed journals, conference papers, books, and book chapters. Recently, his publication is recognized as the ESI Highly Cited Papers. He is a recipient of a number of awards, including the Best Conference Paper Award and the 2016 ACM Transactions on Multimedia Computing, Communications and Applications (TOMM) Nicolas D. Georganas Best Paper Award, Research Quality Award, King Saud University, and the Research in Excellence Award from the College of Computer and Information Sciences (CCIS), King Saud University (3 times in a row). He is on the Editorial Boards of the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE NETWORK, the IEEE MULTIMEDIA, the IEEE WIRELESS COMMUNICATIONS, the IEEE ACCESS, the Journal of Network and Computer Applications (Elsevier), the Computers and Electrical Engineering (Elsevier), the Human-Centric Computing and Information Sciences (Springer), the Games for Health Journal, and the International Journal of Multimedia Tools and Applications (Springer). He also serves as a Lead Guest Editor for the IEEE NETWORK.

Min Chen (minchen2012@hust.edu.cn) is a full professor in School of Computer Science and Technology at Huazhong University of Science and Technology (HUST) since Feb. 2012. He is the director of Embedded and Pervasive Computing (EPIC) Lab at HUST. His Google Scholar Citations reached 15,050+ with an h-index of 60 and i10-index of 188. His top paper was cited 1750+ times. He was selected as Highly Cited Research at 2018. He got IEEE Communications Society Fred W. Ellersick Prize in 2017. His research focuses on cognitive computing, 5G Networks, embedded computing, wearable computing, big data analytics, robotics, machine learning, deep learning, emotion detection, IoT sensing, and mobile edge computing, etc.

Abdulhameed Alelaiwi (aalelaiwi@ksu.edu.sa) is currently an Associate Professor of Software Engineering Department in the College of Computer and Information Sciences, King Saud University (KSU), Riyadh, Saudi Arabia. He received his PhD degree in Software Engineering from the College of Engineering, Florida Institute of Technology-Melbourne, USA in 2002. He is currently serving as the Vice Dean of Research Chairs Program at KSU. He has published over 70+ research papers in the ISI-Indexed journals of international repute. His research interest includes cloud computing, multimedia, Internet of things, Big data, and mobile cloud.

Muneer Al-Hammadi (eng.muneer2008@gmail.com) is a Researcher and a Ph.D. candidate in the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. His research interests include image and video processing, and deep learning