# Deep Emotion Recognition in Dynamic Data using Facial, Speech and Textual Cues: A Survey

Tao Zhang [1,1] and Zhenhua Tan [2]

[1]Software College
[2]Affiliation not available

November 8, 2023



**Abstract**

With the development of social media and human-computer interaction, video has become one of the most common data formats. As a research hotspot, emotion recognition system is essential to serve people by perceiving people's emotional state in videos. In recent years, a large number of studies focus on tackling the issue of emotion recognition based on three most common modalities in videos, that is, face, speech and text. The focus of this paper is to sort out the relevant studies of emotion recognition using facial, speech and textual cues due to the lack of review papers concentrating on the three modalities. On the other hand, because of the effective leverage of deep learning techniques to learn latent representation for emotion recognition, this paper focuses on the emotion recognition method based on deep learning techniques. In this paper, we firstly introduce widely accepted emotion models for the purpose of interpreting the definition of emotion. Then we introduce the state-of-the-art for emotion recognition based on unimodality including facial expression recognition, speech emotion recognition and textual emotion recognition. For multimodal emotion recognition, we summarize the feature-level and decision-level fusion methods in detail. In addition, the description of relevant benchmark datasets, the definition of metrics and the performance of the state-of-the-art in recent years are also outlined for the convenience of readers to find out the current research progress. Ultimately, we explore some potential research challenges and opportunities to give researchers reference for the enrichment of emotion recognition-related researches.

# Deep Emotion Recognition using Facial, Speech and Textual Cues: A Survey

Tao Zhang, and Zhenhua Tan*, *Member, IEEE*

**Abstract**—With the development of social media and human-computer interaction, it is essential to serve people by perceiving people's emotional state in videos. In recent years, a large number of studies tackle the issue of emotion recognition based on three most common modalities in videos, that is, face, speech and text. The focus of this paper is to sort out the relevant studies of emotion recognition using facial, speech and textual cues based on deep learning techniques due to the lack of review papers concentrating on the three modalities. In this paper, we firstly introduce widely accepted emotion models for the purpose of interpreting the definition of emotion. Then we introduce the state-of-the-art for emotion recognition based on unimodality including facial expression recognition, speech emotion recognition and textual emotion recognition. For multimodal emotion recognition, we summarize the feature-level and decision-level fusion methods in detail. In addition, the description of relevant benchmark datasets, the definition of metrics and the performance of the state-of-the-art in recent years are also outlined for the convenience of readers to find out the current research progress. Ultimately, we explore some potential research challenges and opportunities to give researchers reference for the enrichment of emotion recognition-related researches.

**Index Terms**—Emotion recognition, information fusion, deep learning, multimodality.

## 1 Introduction

The past decade has seen the rapid development of emotion recognition in many human-machine interaction and social media applications. Facing external stimulus, human nervous system generates corresponding subjective attitude and expresses emotions via multiple accesses, including face, voice, speech, gait, gesture, and physiological signals such as electroencephalogram(EEG), electrocardiogram(ECG) etc. Emotions play a crucial rule in life, and significantly effect the society behavior and decision-making of human-beings. Given the scene of driving, the emotional states of drivers have significant impact on traffic safety, which means that machines in vehicle need to monitoring emotion states of drivers and give feedback in time. Thus, machine is required to have the ability of recognition, explication and reaction of emotion of human-beings. With regard to social media, a large number of videos are generated by users from all over the world, and uploaded to internet publicly with the characteristics that face, speech and text are the most common modalities. Consequently, it is necessary to develop the research of emotion recognition from these modalities. From the perspective of modalities, researches are divided into unimodal emotion recogition and multimodal emotion recognition. Early researches focus on unimodal emotion recognition such as facial expression recognition (FER), speech emotion recognition (SER) and textual emotion recognition (TER), which attempt to learning emotional features from face, vocals and words of humans, respectively. Some studies also seem other modality as auxiliary to improve the performance of emotion recognition in primary modality during training [1] [2]. Recently, multimodal emotion recognition has gradually been explored and exploited as the complementary among several modalities can significantly improve the accuracy of emotion recognition. The hypothesis is that a large amount of emotional information is dispersed in multiple modalities, so that a suitable learning strategy can fusion information efficiently to obtain better effect. Meanwhile, there are also some efforts to alleviate the inter-subject variations caused by the human attributes such as age, gender and ethnic etc [3]. From another perspective, emotion recognitions in conversations(ERC) and non-conversations are distinguished. Considering a scene that intelligent robots in customer systems are demanded to recognize emotions of customer after customer speaks in a dialogue, two or more parties existed in the dialogue and a party can be influenced by either its own state in the past or the states of other parties. As a result, a series of novel methods are proposed for ERC [4] [5] [6] [7] [8] [9].

Various surveys for emotion recognition have been published in recent years [10] [11] [12] [13] [14] [15] [16] [17] [18]. Li et al. [11] investigate the state-of-the-art for both static and dynamic FERs, and detail the pipelines of FERs in terms of datasets, preprocessing, hand-crafted features, deep learning embedding and comparison of performances, etc. For SERs, [13] [17] also survey fundamental informations of pipelines that are similar with FERs. For TERs, Alswaidan et al. [14] survey the state-of-the-art approaches for TERs but lack an in-depth overview of deep learning-based method for TERs. Deng et al. [10] provide a thorough survey, which systematically review deep learning-based methods for TERs in terms of word embedding, deep learning architecture, training-level approaches and challenges in detail. In addition, several surveys for multimodal emotion recognition are proposed [15] [18] [19]. Mello et al. [19] systematically analysis the meta factors that influence the

• *T. Zhang and Z. Tan are with the College of Software, Northeastern University, Shenyang 110819, China. Z. Tan is the corresponding author. E-mail: zhangtao@stumail.neu.edu.cn, tanzh@mail.neu.edu.cn.*

*Manuscript received April 19, 2005; revised August 26, 2015.*

performance of multimodal emotion recognition. Poria et al. [18] review both emotion recognition and sentiment analysis from unimodality to multimodality. Specifically, they describe available datasets for multimodal emotion recognition and the literature of both unimodal and multimodal emotion recognition focusing on visual, audio and textual modalities. Aside from visual, speech and textual modalities, Jiang et al. [15] consider the real-time mental health monitoring system and take the physiological signal(e.g. EEG, ECG) into consideration.

To our best knowledge, there is a lack of comprehensive surveys about multimodal emotion recognition in recent two years. Meanwhile, the survey focusing on facial, speech and textual modalities that are most common in social media is desperately needed with the rapid development of social network. The purpose of this paper is to fill this gap, which describes the techniques developed in recent years in both unimodality and multimodality. Our contributions are summarized as follows:

1) We propose a systematical review including definition of emotions, research progress of emotion recognition based on deep learning techniques, benchmark datasets, metrics, performances and challenges in future.

2) We conduct a comprehensive introduction of pipelines for unimodal emotion recognition, which includes the techniques of preprocessing, extracting hand-crafted feature and deep feature learning for emotion recognition. For multimodal emotion recognition, we also introduce current research status with respect to feature-level and decision-level fusion.

3) We outline datasets in several modalities as sufficiently as possible, and introduce the attributes of datasets from several perspectives(e.g. modality, year, sample size, the number of subjects, label type, context, language and access). Aside from this, we summarize the performance of methods for emotion recognition on these datasets in recent years and make a clear comparison.

4) We conduct further investigations about existing challenges and oppotunities on the task of multimodal emotion recognition. As a result, some of valuable research challenges and directions are discussed for further research.

The rest of this paper is organized as follows: Section 2 introduces the definition of emotion models. Section 3 describes traditional techniques of preprocessing and handcrafted feature extraction, and the state-of-the-art for emotion recognition in deep learning methods. Section 4 exhaustively introduces dynamic datasets in several dimensions, and summarizes the performance and comparison of methods proposed in recent years on these datasets. Existing challenges and oppotunities are discussed in section 5. Finally, section 6 concludes the paper.

## 2 Emotion Models

Researchers gravitate to the task of emotion recognition over the past two decades owing to the significance of emotion in human nature. To better define and compare the representation of emotion, two kinds of models are proposed: discrete and continuous representation models, which are used commonly in recent researches. Meanwhile, both of discrete and continuous models has obvious advantage and disadvantage. Discrete model limits the representation of emotion in a fixed set, which brings interpretation comprehended intuitively by users, but leads to significant bias while raters annotate erroneously. Continuous model has numerical annotations that are suitable for modelling, but nonintuitive for users. The selection between two kinds of models depends on the actual needs.

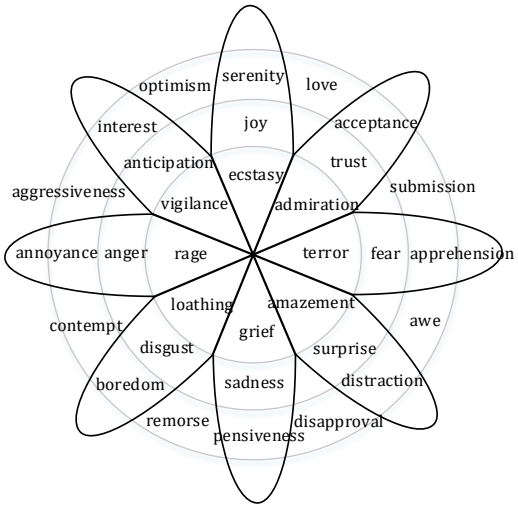



Fig. 1. Plutchik's emotion model [21].

*Discrete representations:* Discrete representations of emotions are widely used due to the intuition and simplicity, which means we can easily have great empathy with the emotion representations. Based on the hypothesis that emotions can be divided into multiple categories, the well-known categories of emotions are proposed by EKman et al. [20], where emotions are separated into six categories: Happiness, Sadness, Fear, Anger, Disgust and Surprise. Furthermore, Plutchik [21] proposes a emotional wheel shown as Fig. 1, which takes the correlation of emotions into consideration and deploys eight emotion categories into four bipolar axes: Joy-Sadness, Fear-Anger, Trust-Digust and Surprise-Anticipation. Each of complex emotions can be represented by the four bipolar axes with different intentions. Note that, one or more discrete emotions mentioned above can occcur simultaneously in a period of dynamic data(e.g. an utterance). Consequently, multi-label emotion recognition are essential when the annotation of samples are discrete.

*Continuous representations:* Aside from discrete representations, emotions can be represented via several numerical dimensions. There are mainly five core dimensions utilized to represent emotions: Pleasure/Valence, Arousal/Activation, Dominance/Power, Anticipation/Expectation and Intensity. Pleasure refers to the degree of positive or negative the expression of person seems. Arousal represents how dynamic or lethargic a person performs. Dominance represents the degree a person feels in control. Strictly speaking, Dominance contains two related concepts: power and control, where power is mainly about internal resources, and control is about the relationship between resources and external factors. There is usu-

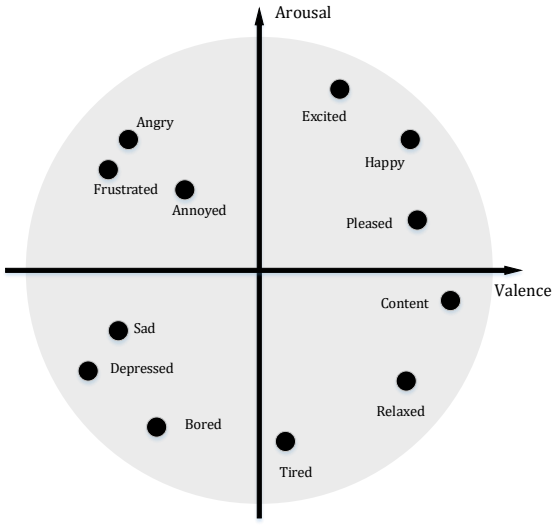



Fig. 2. Circumplex model [22].

ally a composite consideration between two concepts when the annotation of Dominance is processed. Anticipation also subsumes two concepts: expectation and anticipation. Raters take the balance of them into consideration. Intensity is how far a person behaves from the pure retionality. Typically, a person who behaves unemotional is rated as a low score in the dimension of Intensity. Ideally, emotion can be represented to a dimensional vector by these numerical dimensions accurately. In practice, the combination of two or three dimensions are adopted. The most popular emotion representation models are Circumplex model [22] and PAD model [23]. Circumplex model adapts two dimensions to represent emotion shown as Fig. 2: Pleasure and Arousal, which are deemed adequate to represent the most different emotions. Aside from Pleasure and Arousal, PAD model takes also Dominance into consideration.

## 3 DEEP EMOTION RECOGNITION: THE STATE-OF-THE-ART

In this section, we introduce the state-of-the-art methods for emotion recognition in dynamic data using facial, speech, text, and multimodality, respectively. Specifically, we firstly describe the proprecessing techniques and deep feature learning methods that are widely used and achieve good performances in recent years for FERs in dynamic data. Then we introduce SERs in terms of preprocessing, handcrafted feature extraction and deep feature learning. Next, we introduce word embedding methods and deep feature learning methods when text modality is utilized to train model for emotion recognition. Finally, we summarize fusion strategies for emotion recognition in feature-level and decision-level, respectively.

### 3.1 Face

For FER in dynamic data, three macro-manners are utilized to tackle this issue. CNN-RNN: models are set up to recognize emotion in frame-level, and then discriminant results in frame-level are gathered to get ultimate outcome. 3D-CNN: emotion-related latent features are directly extracted via establishing models for spatio-temporal learning. Two-Stream Network: two parallelized accesses are used to learn static and dynamic features, respectively. Here aside from feeding raw video into model for feature learning directly, some of preprocessing techniques, e.g. face detection, alignment, augmentation and normalization, come in handy at times used to enhance the performance of model. We introduce a series of popular techniques with respect to preprocessing and deep feature learning.

#### 3.1.1 Preprocess

The interferences in vision mainly include the noisy information of complex background, variance of illuminations and head pose in unconstrained scenarios [11]. To overcome the interferences of complex background, the usual practice is to detect face region in each frame of video via face detectors, such as Viola-Jones [24]. For the better recognition of emotion, the coordinates of localized landmarks are aligned with face region as the input of training model [25], [26], [27], [28], [29], [30], [31], [32]. Illumination have a adverse effect on the issue of emotion recognition. Thus one of preprocessing steps is to balance the light of face via a series of techniques, e.g. Histogram equalization [33], discrete cosine transform [34] [35], isotropic and anisotropic diffusion [36], difference of Gaussian [37] and homomorphic filtering [38]. There exists a series of pose normalization techniques [39] [40] [41] that yield frontal facial views to overcome head pose problem. It is worth mentioning that pose normalization is essential while emotion recognition is processed in static data(e.g. image) but dynamic data(e.g. video), as the information of emotions can be delivered via head pose. In addition, data augmentation(e.g. cropping, flipping, rotation, shifting, skew, scaling, noise, contrast and color jittering) is overused when static data is preprocessed to overcome the problem of overfitting. When dealing with dynamic data, these data augmentation techniques are also optional to be used in each frame of dynamic data.

#### 3.1.2 Deep Feature Learning

*CNN-RNN*: Frame-level feature learning refers that latent spatial features of emotions are extracted in each selected frame, and then all of the spatial features are aggregated or treated as input of another module for temporal learning. Common structures of frame-level models are Convolutional Neural Network(CNN) - Recurrent Neural Network(RNN) [42] [2] [43] [44] [45] [46] [47] [48] [49], where CNN and RNN are used for spatial and temporal learning in facial modality. For example, Dimitrios Kollias et al. [42] propose a structure that integrates VGG-FACE [50] with GRU [51] to extract temporal and temporal features respectively, where VGG-FACE has been pre-trained with a large dataset for face recognition and achieve excellent performance. The structure of model in [42] is shown as Fig. 3. In addition, the combination of VGG-FACE and LSTM [52] is also adopted in [43] [44]. CNN-RNN structure is more suitable for the feature extraction of macro-expression (longer facial expressions, roughly 24-60 frames) as the features fed into RNN are abstract and global in higher layers of CNN.

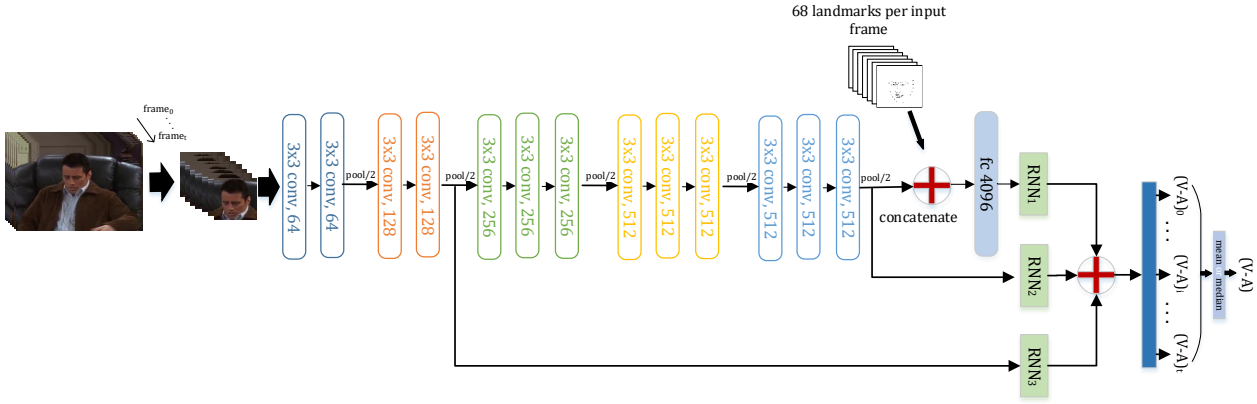



Fig. 3. CNN-RNN structure for emotion recognition proposed in [42].

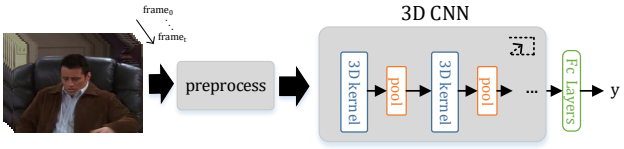



Fig. 4. General 3D-CNN framework.

*3D-CNN*: In recent years, 3D-CNN, as a type of spatio-temporal learning module, has been widely adopted for emotion recognition in videos. The main procedure of 3D-CNN is shown as Fig. 4. Instead of extracting relevant features from each image frame, 3D-CNN directly extracts spatio-temporal features, named C3D, from video via 3D convolutional kernels. 3D-CNN is capable of capture both macro- and micro-expression(roughly 2-10 frames) as 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together which starts from the lowest layer. Specifically, Pre-trained 3D-CNN for Human Action [53] is adopted to extract spatio-temporal features of emotions in videos [54] [55]. Furthermore, Pre-trained 3D-CNN for sports [56] are also applied in a series of studies [7] [8] [57] [58] [59] [60]. Resnet-101 [61], as a 3D-CNN architecture pre-trained on the human action video dataset Kinetics [62], is combined with attention mechanism to form the visual models to extract the feature representation of visual stream [63]. 3D-CNN is gradually being well received when emotional features in videos are required to extract.

*Two-Stream Network*: Compared with CNN-RNN and 3D-CNN, Two-Stream Network is a novel and less well studied architecture for emotion recognition in videos. It is mainly composed of two parallel convolutional networks: a spatial network and a temporal network, which process a static image and instantaneous motion information(e.g. optical flow), respectively. Deng et al. propose a Two-Stream Network named MIMAMO [64], which consists of two stages: Two-Stream convolutional Neural Network and Gated Recurrent Unit Network to capture both micro- and macro-motion, respectively. The feature representation of a snippet, i.e. an RGB image and a sequence of images centered in time around the RGB image, is extracted via the Two-Stream convolutional Neural Network. Specifically, In the temporal network of MIMAMO, Complex Steerable Pyramid [65] is applied to obtain the phase difference between two consecutive facial frames to replace optical flow. These hand-crafted features are fed into CNN to extract latent features. Similarly, the pretrained ResNet50 [66] pretrained on the VGGFace2 face recognition dataset [67] is utilized to extract features from the centering image in the spatial network of MIMAMO. MIMAMO is shown as Fig. 5. Moreover, Pan et al. [68] use two CNNs to extract features from a RGB image and optical flow features in the Two-Stream Network. Feng et al. [69] fed a RGB frame and the hand-crafted feature, i.e. LBP-TOP along the x-t and y-t axes, into two CNNs.

## 3.2 Speech

Speech can effectively deliver emotion information in social activities. In general, the pipeline of SERs mainly contains three steps: preprocessing, extracting hand-crafted features and deep feature leaning. Next, we introduce these steps in more detail.

### 3.2.1 Preprocess

The first step after collecting data is to improve the quality of data via using some of preprocessing techniques for more accurate SER. We describe preprocessing in terms of pre-emphasis, framing, windowing and Voice activity detection.

*Preemphasis:* the average power spectrum of speech signal is influenced by glottic, nose and mouth, which cause the attenuation of power in high-frequency band. The purpose of preemphasis is to compensate high-frequency energy to make flat spectrum over the whole frequency band. Typically, The z-transfer function of high-pass filter for preemphasis is defined as

$$H(z) = 1 - \mu z^{-1} \tag{1}$$

where $\mu \in [0.9, 1.0]$ is the preemphasis coefficient.

*Framing:* framing is to partition speech signal into fixed length segments because of the short-time stationary of speech, which means that speech remains invariant for a sufficiently short period. Besides, there is usually 30% to 50% overlap between two adjacent frames to preserve

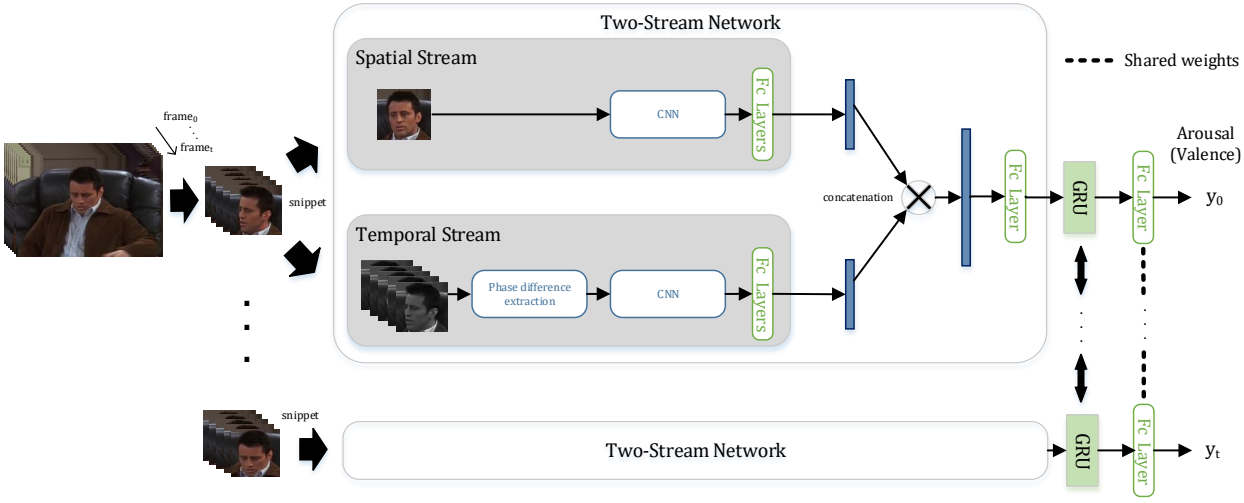



Fig. 5. Two-stream network: MIMAMO proposed in [64].

the information of inter-frames. Consequently, task-related features can be extracted from these quasi-stationary frames for SER.

*Windowing:* the step of framing is processing via a windowing function timing speech signal. Hence, a suitable windowing function need to be allocated. The most overused windowing function are Hamming window ,which is defined as

$$w(n) = 0.54 - 0.46cos(\frac{2\pi n}{N-1}),\ 0 \leq n \leq N-1 \quad (2)$$

where $M$ represents the window size and $w(n)$ represents a frame. Hamming window has the ability to alleviate the effects of leakages that occurs during Fourier Transform caused by discontinuities at the edge of signals. There are also other optional windowing functions, e.g., rectangular window, hanning window, etc.

*Voice endpoint detection:* voice endpoint detection refers to distinguish voice data from unvoice data and noise. Voice data is generated with the vibration of vocal folds that creates periodic excitation to the vocal tract during the pronunciation of phonemes. When air passes through a contriction in the vocal tract, aperiodic excitations are established, which induce transient and turbulent noises, i.e., unvoice data. Typically, zero crossing rate [70], as the rate that the sign of signal changes within a time frame, is a frequently-used method to detect voice endpoint.

### 3.2.2 Hand-crafted Feature

*Prosodic features:* prosodic features have the ability to deliver the significantly distinctive properties of emotions for SER. When people have different emotions, they will have different intonation and rhythm. For example, the emotion of surprise may produces a rising intonation. In order to represent emotions, fundamental frequency($F_0$), energy and duration are widely used in general. Specially, the change of $F_0$ is an essential feature to represent rhythmic and tonal characteristics. For instance, $F_0$ contour usually increases when an emotion of joy is expressed. Other statistics of $F_0$, such as the mean of $F_0$, can also be utilized to represent emotions. Energy, also known as intensity, reflects the amplitude of speech. High arousal emotion is generally accompanied by increased energy, while low arousal emotion is opposite [71]. Emotion status are also represented via the duration-related features, such as the duration of voice, unvoice and silence.

*Voice quality features:* Harmonics-to-Noise ratio(HNR), jitter and shimmer are mostly representable features of voice quality. HNR is the ratio of harmonic components to noise components in speech. Note that the noise here is not ambient noise, but glottic noise caused by incomplete glottic closure. Jitter is a physical properties that describes the change of basic frequency of voice between adjacent vibratory cycles. It mainly reflects the degree of roughness and hoarseness of voice. Shimmer measures the change of amplitude of voice between adjacent vibratory cycles, mainly reflecting the degree of hoarseness.

*Spectral features:* The representation of shape of voice trace is processed in some of sound-related tasks( e.g. automatic speech recognition). The most commonly used representation is spectrum, which transforms speech signal from time domain into frequency domain. Recent research prove that deep learning models can effectively learn emotion representation from spectral features resulting in accurate prediction of emotion categories.

Mel Frequency Cepstral Coefficients (MFCC) is the most frequently-used feature for SERs, which represents shotterm power spectrum of signal. The main procedure is : after preprocessing, short time discrete Fourier transform is applied to speech signal. Then, Mel filter bank is utilized to calculate subband energies in frequency domain, which is followed by the logarithmic computation. Finally, inverse Fourier transform is processed on it. There are a series of spectral features that can be utilized for SERs. Different from MFCC, Gammatone Frequency Cepstral Coefficients (GFCC) applies Gammatone filter-bank to the power spectrum. Besides, Linear Prediction Cepstral Coefficients(LPCC), Log-Frequency Power Coefficients (LFPC), etc are also considerable.

#### 3.2.3 Deep feature Learning

In order to learn emotional representation accurately in speech, deep learning techniques are widely used in a speech emotion recognition system. Typically, on the basis of hand-crafted low-level descriptor, spatio-temporal learning modules are the most commonly used structures to extract deeper and high-level features in latent space to express emotions shown as Fig. 6, which have been proved to be significantly effective for SERs.

Spatio-temporal based sequence-to-one model is suitable for SERs. Various studies model speech signal in segment-level or chunk level to learn emotion representation, and then combine the output of subsequences in segment or chunk level into a sentence-level representation via various strategies(e.g. RNN, Attention) to predict emotions.

Segment-level SERs set the step size of segments as a fixed parameter resulting in changeable number of segments in a utterance, of which the length is arbitrary. Zhang et al. [72] propose a structure that adapts LSTM model to capture utterance-level representation of emotion on the basis of segmeng-level CNN features. Mustaqeem et al. [73] propose CNN+Attention to extract statio-temporal features for SERs. Other studies also achieve satisfying performance for SERs [74] [75]. In addition, chunk-level SERs product a fixed number of segments regardless of the duration of speech signals by changing the step size of the chunks along with the duration of signals. Lin et al. [76] propose a flexible framework, which is capable of tackling several speech-based sequence-to-one tasks(e.g. SER, speaker recognition). They take advantage of varying step size of chunks to extract a fixed number of chunks with a fixed size from varied duration of speech signal without the preprocess such as cropping or zero padding, and the effectiveness of temporal aggregating models that are built via this framework is proved in terms of robustness, accuracy and computational efficiency.

### 3.3 Text

#### 3.3.1 Word Embedding

*Typical word embedding:* Traditional one-hot encoding of word has the disadvantage of high dimension and lack of contextual information. Word embedding models considering syntactic context aims to embed words into low-dimensional space to overcome the drawback of sparsity in traditional models, and has been widely used for natural language processing tasks(NLP). Typical embedding models include Word2Vec [77] [78], GloVe [79], ELMo [80], BERT [81], etc., which are trained on a large number of unlabeled textual data. The former two are trained based on the hypothesis that co-occured words are similar in semantic criteria, and each word has a unique representation. Nevertheless, they are not capable of dealing with the problem brought by polysemy and antonym. In other words, a word may be represented in different context, and antonyms are universally close in embedding space [82]. In recent years, pre-trained language models such as ELMo and BERT are widely adopted in NLP tasks and significantly boost performances. These language models dynamically generate embedding word vectors according to the current context, which make up for the disadvantages existed in the former two models. Other models such as GPT, MASS [83], XLNet [84] are also proposed and focused.

*Emotional word embedding:* Aside from typical word embedding models, emotional word embedding models focusing on embedding words with emotional information are proposed and applied to emotion-related tasks such as emotion recognition and sentiment analysis. There are two kinds of emotional embedding models: word-level and sentence-level embedding models. Emo2Vec [85] is a word-level emotional representation trained with six different emotion-related tasks. DeepMoji [86] is a sentence-level emotional representation that is built by a biLSTM model with attention on a 1246 million tweet corpus, which contains abandant information of emojis(an expression of emotions). Winata et al. [87] analysis and compare both typical and emotional word embedding models in detail, and prove that DeepMoji outperforms other word embedding models by a large margin interpreted as the corpus utilized to train DeepMoji is compatible with emotion-related tasks.

#### 3.3.2 Deep Feature Learning

One of the most common scenario of TERs is dialogue systems. In conversation, semantic information is an important expression of emotions. An appropriate semantic analysis benefits the prediction of emotions. Emotion recognition in conversation has become a hot topic in recent years, and there are tricky issues brought from conversation, that is the contextual dependency of dyadic or multi-party. To tackle this issue, recent studies focus on proposing a series of novel structures to learning contextual representation of emotions, and obtain satisfactory results. Deng et al. [88] propose to integrate general sentence representation and emotional feature representation in sentence-level for further context-level learning. GRU network is utilized to learn contextual information from previous sentences. In addition, Emotion correlations are considered and an emotion correlation learner based on BiGRU is proposed to prediction multi-label emotion. Jiao et al. [89] utilize BiGRU to learn feature and contextual representation of historical utterances as memories, and propose an Attention GRU, of which the state is update by Attention, to predict emotions. Distinguishing from purely supervised learning from a large number of annotated data, Hazarika et al. [90] recognize emotion in conversation from another perspective that transfer the parameters of pre-trained dialogue model on multi-turn conversations to a conversational emotion classifier and achieve significant improvement in terms of performance and robustness. Shen et al. [4] propose an all-in-one XLNet model with enhanced memory to store longer historical context and dialog-aware self-attention to deal with the multi-party structures. How to effectively capture the contextual dependency produced by inter- and intra-party is still demanded to be further explored in future.

### 3.4 Multimodality

Emotions can be expressed via more than one modality of data in social life. One can perceive real-time emotion of another using the fusion of reference evidence from facial, speech and textual information, which are regarded as the effective and acceptable method to improve the acuracy of

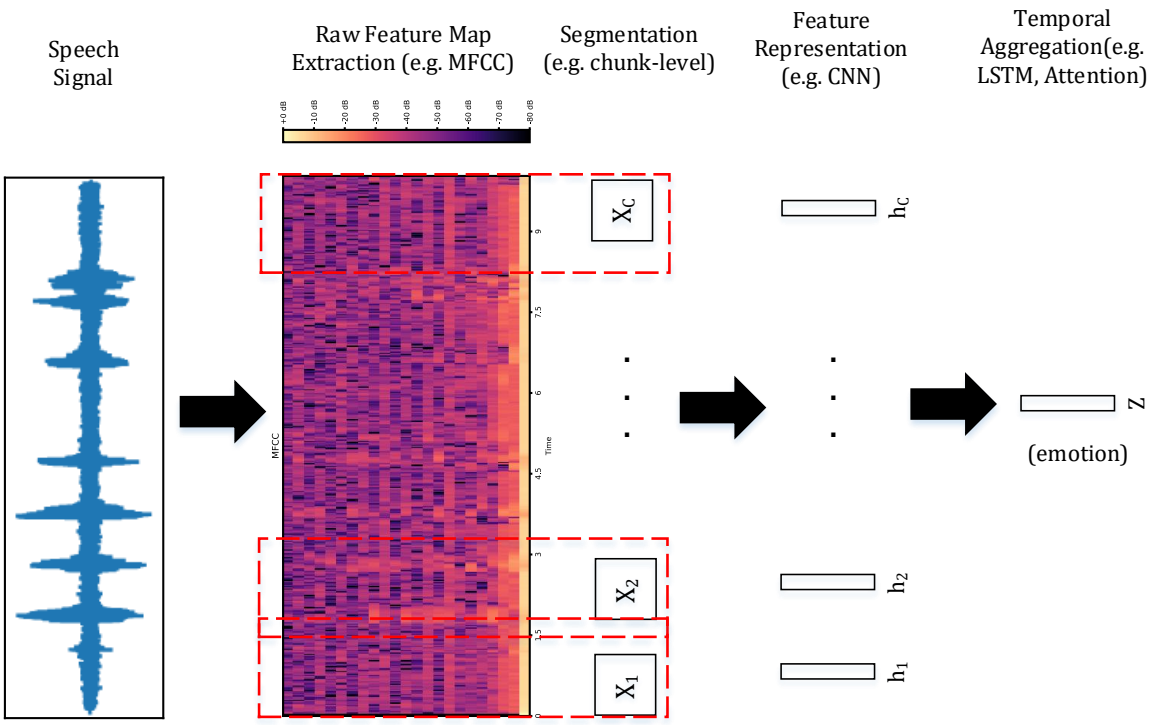
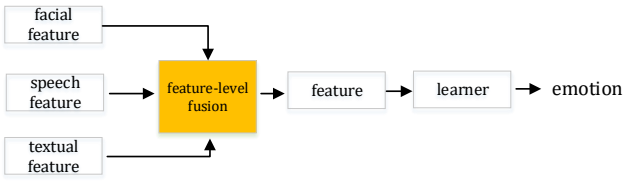
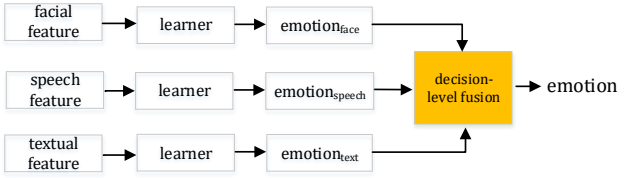



Fig. 6. General framework for speech emotion recognition.

emotion recognition. Recent studies emphasis the utilization of multi-modal data and propose a series of novel structures and proves the advantages of fusion of multimodality. There are mainly two types of multimodal fusion methods shown as Fig. 7: feature-level and decision-level fusion. Feature-level fusion refers that feature representation of emotion from each modality are fused before emotion are inferred. In contrast, Decision-level refers that decision of emotion inference is made in each modality, and then the decision in each modality are fused later. We introduce these two fusion methods emphatically.




(a) Feature-level fusion.




(b) Decision-level fusion.

Fig. 7. Demonstration of fusion in feature-level and decision-level.

### 3.4.1 *Feature-level Fusion*

*Concatenation:* The most widely used strategy of fusion of multi-modalities is to directly concatenate emotion representation generated from each modality. A large number of studies propose novel fusion strategies to tackle multimodal data based on concatenation [8] [57] [58] [63] [93] [94] [95] [9]. Typically, Liang et al. [93] take reconstruction loss and classification loss into consideration, where each modality is constructed via a pair of encoder and decoder, and all of latent representations of different modalities are concatenated as the input of emotion classifier. Zhao et al. [63] utilize temporal attention based 2D ResNet-18 and spatial-channel-temporal attention based 3D ResNet-101 to extract emotion representations from speech and facial data, respectively, and then concatenate emotion representations as the input of fully connected layer for prediction. Hossain et al. [57] concatenate emotion representations extracted from speech and facial modalities as the input of extreme learning machines for emotion recognition. In [8], the concatenation of features of multi-modalities are deemed as the input of GRU to learn global state and personal state with respect to global and personal context in a conversation, respectively.

With the development of Attention [96], one of popular fusion strategy of emotion representations in feature-level is to combine the primitive concatenation strategy with attention, and achieve promising performance for emotion recognition [97] [98] [99]. Mittal et al. [98] integrate fusion layer into memory fusion network [100] to fuse input modalities and classify emotions. Tzirakis et al. [97] test a variety of attention-mechanism based methods, including simple concatenation, self-attention, hierarchical attention, residual self-attention and cross-modal hierarchical self-attention, and prove the effectiveness of cross-modal hierarchical self-attention for fusion of emotion representations. Lian et al. [99] utilize single-modal and cross-modal Transformer to

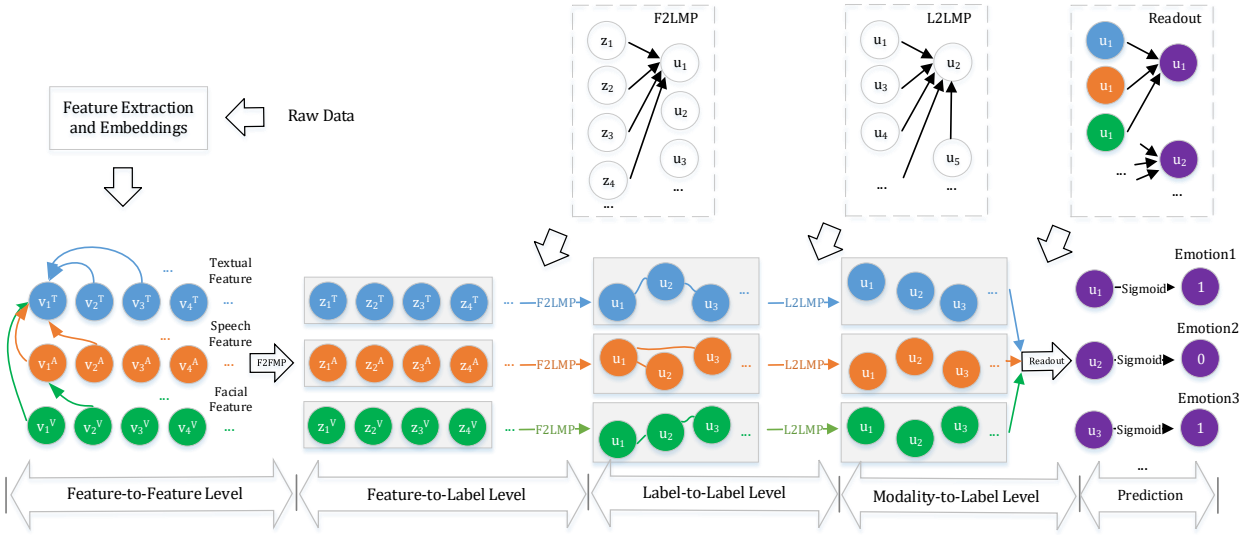



Fig. 8. Heterogeneous Hierarchical Message Passing Network(HHMPN) proposed in [91].

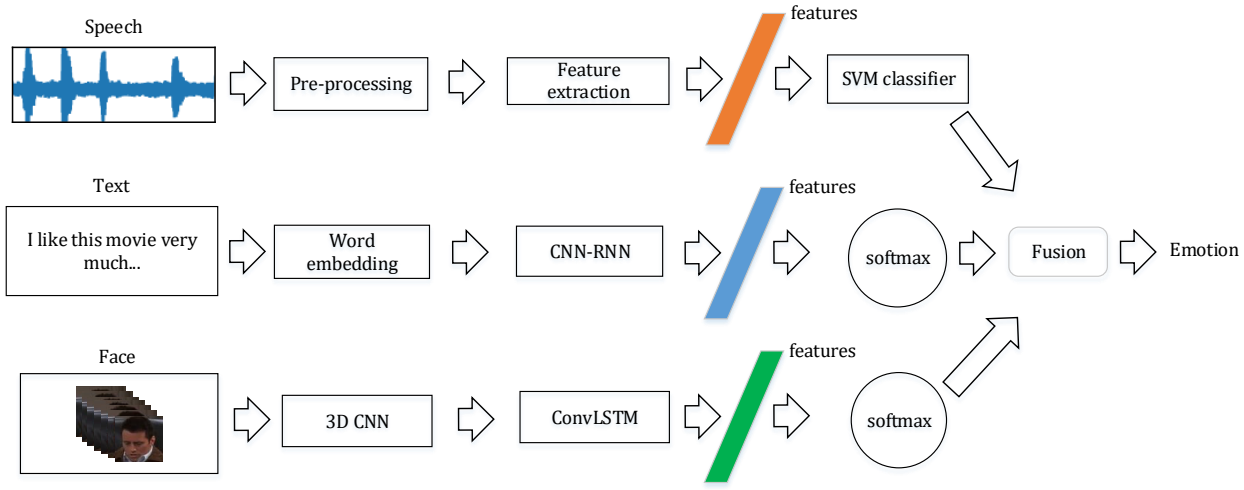



Fig. 9. Model [92] that fuses inference from each modality in decison-level.

extract context-independent utterance-level features, and then propose Audio-Text-Speaker Fusion component for multimodal fusion(ATS-Fusion) and a Multi-Head Attention based bi-directional GRU for contextual feature extraction, where the effectiveness of ATS-Fusion is verified by comparing with simple concatenation.

*Graph network:* Graph neural network(GNN) have been widely applied in a variety of tasks, e.g. information diffusion analysis for online social networks, human action recognition. Recently, GNN has gradually display its capacity for the fusion and feature learning of multi-modalities. To our best knowledge, the latest study of multimodal fusion based on GNN is proposed by [91], which demonstrates the effectiveness of GNN for multi-modal multi-label emotion recognition. In [91], Zhang et al. propose a novel emotion learning structure based on GNN, named as Heterogeneous Hierarchical Message Passing Network(HHMPN). HHMPN is mainly composed of four components shown as Fig. 8. One of components is Feature-to-Feature Level, where each extracted feature of each modality as a node in the graph and messages are passed from a feature node to another one. There are other components , such as Feature-to-Label Level, Label-to-Label Level and Modality-to-Label Level, which work to pass messages among different type of nodes. Benssassi et al. [101] synthesis the representation of neural synchrony graph from facial and speech features that are extracted via spiking neural networks(SNN). The neural synchrony graph is deemed as the input of Graph Convolutional Network(GCN) for emotion recognition.

Some atypical fusion methods are also proposed [43] [102]. Nie et al. [43] fuse speech and facial features by factorized bilinear pooling operation [103]. The process is defined as follows: speech and facial features are fed into the pipeline including fully connected layers, element-wise multiplication, dropout, sum-pooling, L2 normalization to generate high-level descriptor. Wu et al. [102] propose two-

stage fuzzy fusion strategy that combining the Canonical Correlation Analysis(CCA) and Fuzzy Broad Learning System(FBLS) [104] to deal with the imbalanced contributions of each modality and the correlation and difference between multi-modal features. More novel fusion methods are expected to be proposed in forthcoming future.

#### 3.4.2 Decision-level Fusion

Decision-level fusion allocates multiple models considering facial, speech and textual features for emotion recognition. The results generated by unimodal emotion recognition model are combined or aggregated for final prediction. This fusion method is lack of the consideration of correlation among modalities, but flexible to allocate the most suitable model for each modality.

Xu et al. [92] propose a module named 3DCLS (3D Convolutional-Long Short Term Memory) hybrid model to recognize visual emotions and CNN–RNN hybrid model to recognize text-based emotions, which is shown as Fig. 9. In addition, SVM is utilized to recognize emotion from speech. classification probabilities of each emotion category scored from each classifier are weighted and summed to score final probability distribution of emotions. Dahmane et al. [44] apply sequential temporal CNN and LSTM with an Attention Weighted Average layer and Fisher-Vector encoding-based local and global descriptors to generate features from visual and acoustic modality respectively, and then late-fusion is made in decision-level. Multi-task CNN and SVM are utilized to extract features in speech and facial modality, and the fusion is processed via meta-classifier in [105].

There is relatively less studies of fusion in decision-level than that in feature-level. The reason probably is that the expression of emotions is sparse considering inter-modality, which leads to incorrect prediction results in decision-level. Fusing information earlier in feature-level are more suitable to tackle this issue of sparsity.

## 4 DATASETS, METRICS AND PERFORMANCE

### 4.1 Datasets

Sufficiently considered datasets are essential to comprehensively evaluate the performance of emotion recognition system with respect to multi-modalities. In this section, we introduce a series of available datasets, which are widely used for recent research in both unimodality and multimodality. A summary of datasets is shown in TABLE 1, where datasets are grouped by the categories of modalities. Here we do not cover image-related datasets, as the primary purpose of the paper is to inform readers on dynamic datasets such as videos.

#### 4.1.1 Multimodality: Face, Speech, Text

**IEMOCAP** [106]: Interactive emotional dyadic motion capture database(IEMOCAP) is collected by the Speech Analysis and Interpretation Laboratory(SAIL) at the University of Southern California (USC). IEMOCAP totally contains 10039 turns of 10 speakers with an average duration of 4.5 seconds in both scripted and spontaneous conversations. The corpus contains approximately 12 h of data, which marks on facial expressions, hand movements, audio and transcriptions with both categorical(happiness, sadness, anger, surprise, fear, disgust, frustration, excited and neutral) and continuous(activation, arousal and dominance) annotations.

**AVEC** [107]: Audio-Visual Emotion recognition Challenge(AVEC), which uses the Solid-SAL part of SEMAINE [130] database, contains 5816 utterances of 150 participants in conversations with four real valued affective attributes: valence, arousal, expectancy, and power.

**CMU-MOSEI** [108]: CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) is collected from social multimedia. CUM-MOSEI contains 23453 annotated sentences from 1000 speakers, which comprise six categorical emotions: angry, disgust, fear, happy, sad and surprise.

**OMG** [109]: The One-Minute Gradual-Emotional Behavior dataset(OMG) contains in-the-wild videos collected from social multimedia YouTube. OMG totally contains 7371 annotated utterances based on monologued scenarios with both categorical(anger, disgust, fear, happiness, sadness, surprise and neutra) and continuous(valence and arousal) annotations.

**MELD** [110]: Multimodal EmotionLines Dataset (MELD) is an extension and enhancement of EmotionLines dataset [131]. MELD has more than 1400 dialogues and 13000 utterances of 304 speakers from Friends TV series. Here multi-speakers participate in a conversation. Utterances are annotated in seven categorical emotions: anger, disgust, sadness, joy, neutral, surprise and fear.

**NNIME** [111]: The NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus (NNIME) adapts the use of dyadic interactions for natural elicitation of affective behaviors where participants are with prior real-life experiences in professional acting performances. NNIME contains 6701 utterances of 44 participants with both categorical(anger, frustration, sadness, surprise, neutral, and happiness) and continuous emotions(valence and arousal).

**NEMu** [91]: NEMu is a multi-label multi-modal dataset collected from NetEase Cloud Music in Chinese, where contains 18907 samples and each sample is composed of multiple modalities: lyric, comments, audio and images. Twelve discrete emotion labels are annotated on this dataset.

**SEWA** [112]: The Sentiment Analysis in the Wild (SEWA) are collected from web-cameras and microphones, which records the reactions of participants while watching and discussing adverts, respectively. SEWA consists of 1990 video clips with 2562 utterances of 398 speakers from different cultures, including chinese, english, german, greek, hungarian and serbian. Continuous attributes of emotions are annotated: valence and arousal.

#### 4.1.2 Multimodality: Face, Speech

**eNTERFACE'05** [113]: eNTERFACE'05 is an audio-visual emotion database experienced with pre-defined situations and reactions. eNTERFACE'05 consists of a total of 1166 video sequences of 43 different speakers who are asked to speak in English in the experiment despite various countries. Six basic emotions of video sequences are annotated, including anger, disgust, fear, happiness, sadness and surprise.

**SAVEE** [114]: Surrey Audio-Visual Expressed Emotion (SAVEE) database is recorded in a visual lab from 4 actors in seven emotions: anger, disgust, fear, happiness, sadness,

TABLE 1
Summary of datasets. F=face S=speech, T=text, VAL=valence, ARO=arousal, EXC=exceptancy, DOM=dominance, POW=power. P=posed, N=natural.

| Modality | Dataset | Year | Samples | Classes | Label | dialog | language | Type | Access |
|---|---|---|---|---|---|---|---|---|---|
| F, S, T | IEMOCAP [106] | 2008 | 10,039 samples of 10 subjects | categorical:nine continuous:VAL,ARO,DOM | mono | yes | English | P, N | https://sail.usc.edu/iemocap/ |
| | AVEC [107] | 2012 | 5816 samples of 150 subjects | continuous:VAL,ARO,EXC,POW | mono | yes | English | P | https://semaine-db.eu/ |
| | CMU-MOSEI [108] | 2018 | 23453 samples of 1000 subjects | categorical:six | multi | no | English | N | https://github.com/A2Zadeh/CMU-MultimodalSDK |
| | OMG [109] | 2018 | 7371 samples | categorical:seven continuous:VAL,ARO | mono | no | English | N | https://github.com/knowledgetechnologyuhh/OMGEmotionChallenge |
| | MELD [110] | 2019 | 13708 samples of 304 subjects | categorical:seven | mono | yes | English | N | https://github.com/SenticNet/MELD |
| | NNIME [111] | 2017 | 6701 samples of 44 subjects | categorical:six continuous:VAL,ARO | mono | yes | Chinese | N | http://nnime.ee.nthu.edu.tw/ |
| | NEMu [91] | 2021 | 18907 samples | categorical:twelve | multi | no | Chinese | N | https://github.com/MANLP-suda/HHMPN |
| | SEWA [112] | 2021 | 2562 samples of 398 subjects | continuous:VAL,ARO | mono | yes | multi | N | http://db.sewaproject.eu |
| F, S | eNTERFACE'05 [113] | 2006 | 1166 samples of 43 subjects | categorical:six | mono | no | English | P | http://www.enterface.net/enterface05 |
| | SAVEE [114] | 2008 | 480 samples of 4 subjects | categorical:seven | mono | no | English | P | http://kahlan.eps.surrey.ac.uk/savee/ |
| | AFEW [115] | 2012 | 957 samples of 330 subjects | categorical:seven | mono | no | English | N | https://cs.anu.edu.au/few/ |
| | RECOLA [116] | 2013 | 1308 samples of 23 subjects | continuous:VAL,ARO | mono | yes | French | N | http://diuf.unifr.ch/diva/recola |
| | VideoEmotion 8 [117] | 2014 | 1101 samples | categorical:eight | mono | no | English | N | www.yugangjiang.info/research/VideoEmotions/ |
| | CREMA-D [118] | 2014 | 7442 samples of 91 subjects | categorical:six | mono | no | English | P | https://github.com/CheyneyComputerScience/CREMA-D |
| | LIRIS-ACCEDE [119] | 2015 | 9800 samples | continuous:VAL,ARO | mono | no | Mainly english | N | http://liris-accede.ec-lyon.fr/ |
| | BAUM-1 [120] | 2017 | 1184 samples of 31 subjects | categorical:nine | mono | no | Turkish | N | http://baum1.bahcesehir.edu.tr/ |
| | MSP-Improv [121] | 2017 | 8,438 samples of 12 subjects | categorical:four | mono | no | English | P, N | https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Improv.html |
| | Ekman-6 [122] | 2018 | 1637 samples | categorical:six | mono | no | English | N | https://github.com/kittenish/Frame-Transformer-Network |
| | RAVDESS [123] | 2018 | 7356 samples of 24 subjects | categorical:eight | mono | no | English | P | https://doi.org/10.5281/zenodo.1188976 |
| S, T | EMO-DB [124] | 2005 | 800 samples of 10 subjects | categorical:seven | mono | no | German | P | http://www.expressive-speech.net/emodb/ |
| | MSP-Podcast [125] | 2017 | 62,140 samples | categorical:eight continuous:VAL,ARO,DOM | mono | no | English | N | https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html |
| S | TESS [126] | 2020 | 2800 samples | categorical:seven | mono | no | English | P | https://doi.org/10.5683/SP2/E8H2MF |
| T | Ren-CECps [127] | 2010 | 35096 samples | categorical:eight | multi | no | Chinese | # | https://github.com/KGBUSH/Ren_CECps-Dictionary |
| | EMOBANK [128] | 2017 | 10548 samples | attribute:VAL,ARO,DOM | mono | no | English | # | https://github.com/JULIELab/EmoBank |
| | DailyDialog [129] | 2017 | 102979 samples | categorical:seven | mono | yes | English | # | http://yanran.li/dailydialog |
| | SemEval2018 [130] | 2018 | 10983 samples | categorical:eleven | multi | no | English | # | https://competitions.codalab.org/competitions/17751 |

surprise and neutral. 480 utterances are collected and annotated in SAVEE.

**AFEW** [115]: Acted Facial Expressions in the Wild (AFEW) is a dynamic database collected from movies where a scene can contain multiple subjects. AFEW is closed to real world environment, which contains 330 subjects and 957 video clips in seven emotions: anger, disgust, fear, sadness, happiness, neutral and surprise.

**RECOLA** [116]: Spontaneous collaborative and affective interactions(RECOLA) corpus records multimodal conversation of a pair of participants continuously and synchronously in dyads when participants work to tackle a task collaboratively over video conference. 1308 utterances of 23 participants are publicly avaliable and annotated in continuous attributes of emotions: valence and arousal.

**VideoEmotion8** [117]: VideoEmotion8 collects user-generated videos from popular video-sharing websites YouTube and Flickr. 1101 videos are downloaded and annotated in eight emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. A minimum number of 100 videos per category and an average duration of 107 seconds are satisfied.

**CREMA-D** [118]: Crowd-Sourced Emotional Multimodal Actors Dataset(CREMA-D) induces 91 actors diverse ethnic backgrounds to repeat 12 sentences multiples times in six categorical emotions(anger, fear, disgust, neutral, happy, sad). 7442 utterances are recorded in CREMA-D.

**LIRIS-ACCEDE** [119]: LIRIS-ACCEDE consists of 9800 video excerpts with a large content diversity and each is from 8 to 12 seconds long. The valence and arousal attributes are annotated as continuous labels of emotion.

**BAUM-1** [120]: BAUM-1 presents audio-visual affective face in a spontaneous manner. Subjects are shown a sequence of images and short video clips, which can evoke a set of emotions, and then subjects express ideas and feeling in an unscripted in Turkish. BAUM-1 encompasses 1184 video clips of 31 subjects in nine categorical emotions(happiness, sadness, anger, fear, disgust, surprise, neutral, boredom and contempt).

**MSP-Improv** [121]: Conversational dyadic improvisations are processed to capture naturalistic emotions while partially controlling for lexical content. 652 target sentences

are improvised and 620 target sentences are reading by actors. The number of remainder of improvised sentences is 4381. Natural interaction (interactions between recordings of the scenarios) contains 2785 sentences. Totally, 8348 sentences are recorded in MSP-Improv.

**Ekman-6** [122]: Ekman-6 comprises 1,637 videos collected from Youtube and Flickr. The average duration is 112 seconds. The videos are labeled in six categorical emotions: anger, disgust, fear, sadness, joy and surprise.

**RAVDESS** [123]: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) comprises 7356 recordings of 24 professional actors who reading a sentence in different emotions. Eight categorical emotions are considered for speaking: neutral, calm, happy, sad, angry, fearful, surprise, and disgust. In addition, song is with six emotions(neutral, calm, happy, sad, angry, and fearful).

#### 4.1.3 Multimodality: Speech, Text

**EMO-DB** [124]: Ten actors express ten pre-scripted German utterances in seven emotions(neutral, anger, fear, joy, sadness, disgust, boredom). Ultimately, EMO-DB comprises about 800 sentences.

**MSP-Podcast** [125]: MSP-Podcast is an ongoing process, which retrieving speech from existing podcast recordings. There is 62140 speaker turns(100 hours) in both categorical(anger, happiness, sadness, disgust, surprised, fear, contempt and neutral) and continuous(dominance, arousal and valence) emotion labels up to now. The purpose of the corpus is to build a 400 hours corpus with emotionally balanced labels.

#### 4.1.4 Unimodality: Speech

**TESS** [126]: Toronto emotional speech set (TESS) contains 2800 sentences. Two actresses who are recruited from the Toronto are asked to speak pre-defined phrases in each of seven categorical emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral).

#### 4.1.5 Unimodality: Text

**Ren-CECps** [127]: Ren-CECps is a chinese corpus collected from blog texts. The document, paragraph and sentence of Ren-CECps are annotated in each of nine categorical emotions(joy, hate, love, sorrow, anxiety, surprise, anger, expect and neutral). 35096 sentences are annotated in Ren-CECps.

**EMOBANK** [128]: A total of 10548 English sentences with balanced genres in six domains(news, fictions, blogs, essays, letters and travel Guides) are annotated. Three continuous dimensions of emotions are considered and labeled: valence, arousal and dominance.

**DailyDialog** [129]: DailyDialog records 13118 Dialogues each with 7.9 of average Speaker Turns. There are totally 102979 utterances with annotated emotion labels. Utterances contains the main six categorical emotions: anger, disgust, fear, happiness, sadness and surprise. The corpus covers our daily topics of communication with high-quality multi-turn dialogs.

**SemEval2018** [130]: SemEval2018, which consists of 11 emotion labels: anticipation, anger, fear, joy, disgust, love, optimism, sad, surprise, trust and pessimism, is collected from tweets for competition. A total of 10983 sentences are included and annotated.

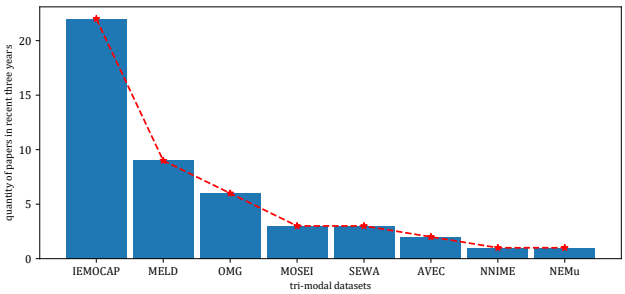



Fig. 10. Distribution of tri-modal datasets.

### 4.2 Evaluation Metrics

Evaluation of models is essential to advance the progress of research. We introduce widely used evaluation metrics to evaluate the performance of models for both categorical and continuous emotion recognition.

#### 4.2.1 Categorical

For categorical emotion recognition, most of the state-of-the-arts utilize $Accuracy$(or called $Recall$) [99] and $F1$ score to evaluate the performance of models. Here we suppose there are $C$ emotion classes in a dataset. $N_c$ represents the number of samples of class $c$, where $c \in \{1, 2, ..., C\}$. For class $c$,

$$Prediction_c = \frac{TP_c}{TP_c + FP_c}. \tag{3}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c}. \tag{4}$$

$$F1_c = \frac{2 * Precision_c * Recall_c}{Precision_c + Recall_c}. \tag{5}$$

where $TP_c$ is the true positive of class $c$, $FP_c$ is the false positive of class $c$, $TN_c$ is the true negative of class $c$, $FN_c$ is the false negative of class $c$. Other metrics are defined as (6)(7)(8)(9).

- $Weighted\ average\ accuracy(ACC)$:

$$ACC = \frac{\sum_{c=1}^{C} N_c * Recall_c}{\sum_{c=1}^{C} N_c} \tag{6}$$

- $Unweighted\ average\ accuracy(uACC)$:

$$uACC = \frac{\sum_{c=1}^{C} Recall_c}{C} \tag{7}$$

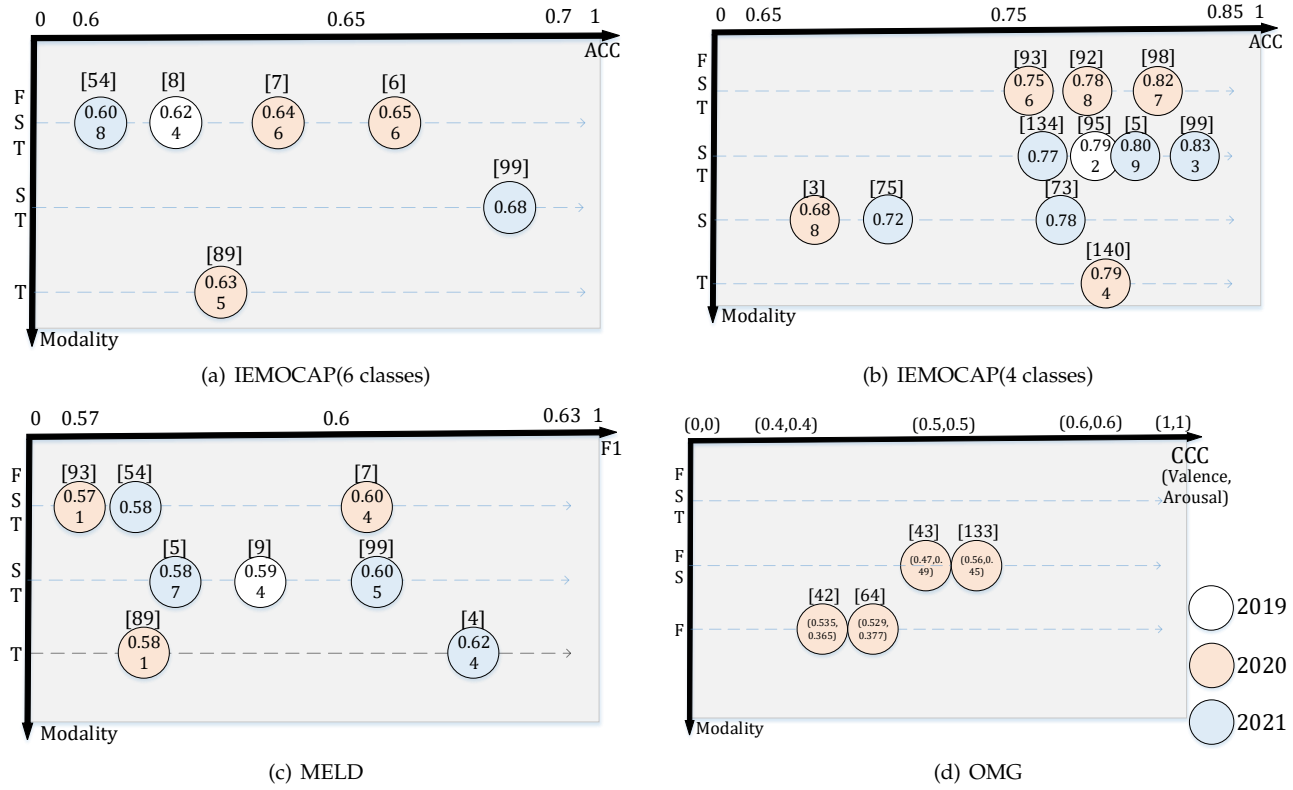
- $Weighted\ average\ F1$:

$$F1 = \frac{\sum_{c=1}^{C} N_c * F1_c}{\sum_{c=1}^{C} N_c} \tag{8}$$

- $Unweighted\ average\ F1(uF1)$:

$$uF1 = \frac{\sum_{c=1}^{C} F1_c}{C} \tag{9}$$

where $N_c$ represents the number of samples of class $c$ in dataset.




(a) IEMOCAP(6 classes) (b) IEMOCAP(4 classes)

(c) MELD (d) OMG

Fig. 11. Performances of models for emotion recognition on tri-modal dataset in recent three years.

#### 4.2.2 *Continuous*

Both the Pearson Correlation Coefficients($PCC$) and Concordance Correlation Coefficient ($CCC$) are widely used to estimate the performance of emotion recognition with continuous annotations. We suppose that $y_i$ represents true value of sample $i$ and $\hat{y}_i$ represents predicted value of sample $i$. The definitions of $PCC$ and $CCC$ are shown as (10)(11).

- Pearson Correlation Coefficients:

$$PCC = \frac{\sigma^2_{y\hat{y}}}{\sigma_y \sigma_{\hat{y}}} \tag{10}$$

where $\sigma^2_{y\hat{y}}$, $\sigma_y$ and $\sigma_{\hat{y}}$ represents covariance between $y$ and $\hat{y}$, standard deviation of $y$ and standard deviation of $\hat{y}$, respectively.

- Concordance Correlation Coefficient:

$$CCC = \frac{2\sigma^2_{y\hat{y}}}{\sigma^2_y + \sigma^2_{\hat{y}} + (\mu_y - \mu_{\hat{y}})^2} \tag{11}$$

where $\sigma^2_{y\hat{y}}$ represents covariance between $y$ and $\hat{y}$, $\sigma^2_y$ represents variance of $y$, $\sigma_{\hat{y}}$ represents variance of $\hat{y}$, $\mu_y$ represents mean value of $y$ and $\mu_{\hat{y}}$ represents mean value of $\hat{y}$.

### 4.3 Performances and discussion

A large number of studies put emphases on the improvement of performance on the task of emotion recognition. TABLE 2 summarizes the state-of-the-art of emotion recognition in multimodal data in terms of modality, year of publication, benchmark datasets, metrics and performance. Similarly, TABLE 3 summarizes the state-of-the-art of emotion recognition in unimodal data. Fig. 10 illustrates the distribution of quantity of papers on each dataset. The statistical results show that IEMOCAP, MELD and OMG rank the top three in terms of frequency of use. Consequently, Fig. 11 displays the performances of models proposed in recent three years on three tri-modal datasets: IEMOCAP, MELD and OMG, where metric is plotted on the horizontal x-axis against modality is plotted on the vertical Y-axis for the convenience that readers can clearly see the research trend for emotion recognition on these datasets. Note that in IEMOCAP, 6 classes(anger, happiness, sadness, neutral, excitement and frustration) of discrete categories are considered in Fig. 11(a) for the convenience of comparison of state-of-the-art frameworks. Meanwhile, there are also some studies that consider 4 classes(anger, happiness, sadness, neutral) and merge happiness and excitement categories into the single happiness category to alleviate the variance caused by ambiguous annotations, which is shown as Fig. 11(b).

As can be seen from Fig. 10, IEMOCAP is the most commonly used multimodal benchmark compared with other datasets. The reason may be that IEMOCAP was available as early as 2008 and has a large number of samples and rich annotations, which makes this dataset popular for emotion recognition in both unimodality and multi-modality. In addition, the dataset is generated in dialogue, thus a large number of ERCs use IEMOCAP as a benchmark. MELD, as another multimodal conversation dataset, was published in 2019 and has the characteristics of a large number of samples and subjects. Therefore, MELD has been gradually

adopted as benchmark in recent years. For non conversational emotion recognition, OMG is an in-the-wild dataset with a large number of samples and participants. MOSEI is the few multi-label and multi-modal dataset, which leads to frequent use in the research of multi-modal and multi-label emotion recognition. Other multimodal datasets are used relatively infrequently in research.

It can be found from TABLE 2 and TABLE 3 that the most commonly used metrics for discrete emotion labels are $ACC$, $uACC$, $F1$ and $uF1$, and that for continuous emotion labels is $CCC$. However, for discrete emotion reccognition, some studies only use part of $ACC$, $uACC$, $F1$ and $uF1$, which makes it difficult for readers to directly compare performances of studies using different metrics. We expect that the future research will comprehensively apply the metrics to the experiment. We can also find that more emotion recognition methods based on multimodal fusion are proposed and achieve good performances. Meanwhile, the trend of research of emotion recognition based on unimodality is still positive, especially SERs and TERs are also research hotspots in recent years. Consequently, more innovative research is expected to be put forward.

## 5 CHALLENGES AND OPPOTUNITIES

### 5.1 Privacy enhancement

With the development of social media and artificial intelligence, many applications aim to recognize emotions of users. this procedure often needs users to transmit data to server, where the transmission is vulnerable to hacking and re-identification. Eavesdroppers can easily obtain sensitive information from eavesdropped data. An improved strategy that protects privacy is to transmit data representation generated on devices to server, and then the data representation is processed on server by presupposition mechanism for further tasks. Nevertheless, sensitive demographic informations are also leakaged through certain technology. Demographic information such as race, age and gender are significant in hiring, policing and credit ratings. Therefore, one primary task of privacy concerns is to eliminate demographic information that existed in the representations of multimodal data, while maintaining the performance on the task of emotion recognition. To our best knowledge, the first study on privacy enhanced emotion recognition in multimodal data is proposed by [94]. We expect more studies focusing on privacy issues occur in future.

### 5.2 Generalization and personalization

It is well-known that emotions are percepted by humanbeing from the whole world in a consistent manner [142]. However, building a generic enough model to recognize emotions in natural scenarios from different persons is difficult. The major causes are the difference of emotion expression of each person and the variety of understanding and display of emotion depending on different situation, interaction partner and even the time of the day [143] [144] [145]. In addition, available in-the-wild datasets usually contains short-time expression of an individual. Only a few frames or tens of seconds are collected that lead to poor personalization.

Recent studies focus on exploring more generalized models based on deep learning techniques, which are trained on a large number of in-the-wild datasets with emotion labels. These studies aim to extract features from different individuals and maximize the generalization on emotion expression recognition. Nevertheless, they perform excellent on training datasets but poorly on real-world problems as the expression of a person maybe extremely personalized and not well represented in training datasets. For example, the expression of happiness of a person could seems like an expression of sadness sometimes. Thus, personalization of models are essential to adapt to new individuals.

To overcome the poor performance produced by personalization of individuals, Transfer learning and lifelong learning techniques are utilized to address this problem. Cross-corpus emotion recognition [146] [147] aims to improve performance of emotion recognition on target domain by transferring knowlegde from source domain with labelled samples to target domain with less or not labelled samples. These models present a significant improvement on target datasets but exist some limitations when applied to real-world scenarios due to the expensive and slow adaptation process. Lifelong learning is considered as an major breakthrough in the fight against the balance between generalization and personalization. Barros et al. [133] adapt the interplay between generalization and personalization by self-organizing mechanisms that interpret emotions from self-organized general emotion representation and personalized emotion recognition. Incoming samples are organized incrementally that means no retrain is required. Lifelong, unsupervised representation learning for both generalized and personalized emotion recognition in multimodality need to be further explored in future.

### 5.3 Unified model

A large number of studies concentrate on proposing novel models for emotion recognition in unimodality, bimodality and multimodality. Despite this, there is still a small amount of unified model simultaneously suitable for each modality and arbitrary combination of modality. The studies focusing on multimodal fusion experiment on multimodal data for the purpose of accurate improvement of emotion recognition, but are short of the empirical evidence to prove the effectiveness of models when some of modalities are unavailable. Mittal et al. [98] propose M3ER that utilizes Modality Check Step to replace unavailable modality with proxy feature and fuses multimodal features by multiplicative fusion module. M3ER is a promising technique but similarly lack of experiments in unimodality and bimodality. [97] [132] evaluate their models in unimodality and multimodality, but lacks the experiments in bimodality. [6] propose a model for real-time emotion detection in conversations and experiment in bimodality and multimodality data but lack of the unimodality. [7] and [91] evaluate the performances of their models using different modality combinations but lack the comparison against baselines in unimodality and bimodality. Liang et al. [93] evaluate their model and empirical analysis is presented in unimodality, bimodality and mulmodality. In short, one of the challenge is to propose unified models with sufficient comparative experiments in unimodality, bimodality and mulmodality.

TABLE 2
Summary of the state-of-the-art in multimodality proposed in recent three years. F=facce, S=speech, T=text.

| Modality | Method | Year | Datasets | Metrics | Performance |
|---|---|---|---|---|---|
| F, S, T | Q. Li et al. [54] | 2021 | IEMOCAP<br>MELD | ACC, F1 | ACC: 0.6084, F1: 0.5988(IEMOCAP, 6 classes)<br>ACC: 0.6081, F1: 0.58(MELD, 7 classes) |
| | D. Zhang et al. [91] | 2021 | CMU-MOSEI<br>NEMu | ACC | 0.459(MOSEI, 6 classes)<br>0.249(NEMu, 12 classes) |
| | P. Tzirakis et al. [97] | 2021 | SEWA | CCC | Valence0.783, Arousal0.690 |
| | T. Mittal et al. [98] | 2020 | IEMOCAP<br>CMU-MOSEI | uACC | 0.827(IEMOCAP, 4 classes)<br>0.89(CMU-MOSEI, 6 classes) |
| | D. Zhang et al. [6] | 2020 | IEMOCAP<br>AVEC | ACC,F1, PCC | ACC: 0.656, F1: 0.653(IEMOCAP, 6 classes),<br>PCC: Valence0.389 Arousal0.39 Expectancy0.403 Power0.387(AVEC) |
| | X. Ju et al. [132] | 2020 | CMU-MOSEI | ACC | 0.494(6 classes) |
| | J. Liang et al. [93] | 2020 | IEMOCAP<br>MELD | uACC, ACC, F1 | uACC: 0.745, ACC: 0.756(IEMOCAP, 4 classes)<br>F1: 0.571(MELD, 7classes) |
| | S. Xing et al. [7] | 2020 | IEMOCAP<br>MELD | ACC, F1 | ACC: 0.646, F1: 0.643(IEMOCAP, 6 classes)<br>F1: 0.6045(MELD, 7 classes) |
| | G. Xu et al. [92] | 2020 | IEMOCAP | ACC | 0.7875(4 classes) |
| | N. Majumder et al. [8] | 2019 | IEMOCAP<br>AVEC | F1, PCC | F1: 0.629 (IEMOCAP, 6 classes)<br>PCC: Valence0.37 Arousal0.6 Expectancy0.37 Power0.41(AVEC) |
| F, S | S. Zhao et al. [63] | 2020 | VideoEmotion8<br>Ekman-6 | ACC | 0.545(VideoEmotion8, 8 classes)<br>0.553(Ekman-6, 6 classes) |
| | E. M. Benssassi et al. [101] | 2020 | eNTERFACE'05<br>RAVDESS | ACC | 0.9682(eNTERFACE'05, 6 classes)<br>0.983(RAVDESS, 6 classes) |
| | M. Dahmane et al. [44] | 2020 | OMG | CCC | Valence0.47 Arousal0.49 |
| | P. Barros et al. [133] | 2020 | OMG | CCC | Valence0.56 Arousal0.45 |
| | W. Nie et al. [43] | 2020 | eNTERFACE'05<br>AFEW | uACC | 0.9707(eNTERFACE'05, 6 classes)<br>0.6355(AFEW, 7 classes) |
| | M. Wu et al. [102] | 2020 | SAVEE<br>eNTERFACE'05<br>AFEW | ACC | 0.9979(SAVEE, 7 classes)<br>0.9082(eNTERFACE'05, 6 classes)<br>0.5028(AFEW, 7 classes) |
| | M. Hao et al. [105] | 2020 | eNTERFACE'05 | ACC | 0.8136(6 classes) |
| | Y. Ma et al. [58] | 2019 | eNTERFACE'05<br>BAUM-1 | ACC | 0.8394(eNTERFACE'05, 6 classes)<br>0.5761(BAUM-1, 6 classes) |
| | M. S. Hossain et al. [57] | 2019 | eNTERFACE'05 | ACC | 0.864(6 classes) |
| S, T | S. Zhou et al. [134] | 2021 | IEMOCAP | uACC, ACC | uACC: 0.77, ACC: 0.766(4 classes) |
| | Z. Lian et al. [99] | 2021 | IEMOCAP<br>MELD | uACC, ACC, uF1, F1 | uACC: 0.833, ACC: 0.836, F1:83.8(IEMOCAP, 4 classes)<br>uACC: 0.676, ACC: 0.68, uF1:0.67, F1:0.675(IEMOCAP, 6 classes)<br>ACC: 0.62, F1: 0.605(MELD, 7 classes) |
| | Z. Lian et al. [5] | 2021 | IEMOCAP<br>MELD | ACC, F1 | ACC: 0.809, F1: 0.8081(IEMOCAP, 4 classes)<br>ACC: 0.6134, F1: 0.5865(MELD, 7 classes) |
| | M. Jaiswal et al. [94] | 2020 | IEMOCAP<br>MSP-Improv<br>MSP-Podcast | uACC | discrete activation0.68, discrete valence0.69(IEMOCAP),<br>discrete activation0.63, discrete valence0.51(MSP-Improv),<br>discrete activation0.70, discrete valence0.56(MSP-Podcast) |
| | Y. Gu et al. [135] | 2019 | IEMOCAP<br>MELD | uACC, uF1, F1 | uACC: 0.516, F1: 50.3(IEMOCAP, 9 classes)<br>uACC:0.794, uF1:75.3(anger), uACC:0.704, uF1:70.1(Joy), uACC:0.657, uF1: 65.4(neutral),<br>uACC:0.84, uF1:79.2(sad), uACC:0.783, uF1: 74.0(surprise)(MELD, Binary classification) |
| | R. Li et al. [95] | 2019 | IEMOCAP | uACC, uF1 | uACC: 0.792, uF1: 0.791(4 classes) |
| | D. Zhang et al. [9] | 2019 | MELD | F1 | 59.4(7 classes) |

### 5.4 Datasets with abundant samples

Deep learning model trained on single dataset usually lacks generalizability when cross-dataset setting is configured. The main reason is the difference of conditional distribution between different datasets caused by inconsistent expression annotations. Because of different collecting conditions and the subjectiveness of annotation, the performance of emotion recognition model cannot keep improving when enlarging the training data by directly merging multiple datasets. Therefore, accurately annotating a large volume of multimodal data with the large variation and complexity of natural scenarios is an obvious impediment to the construction of expression datasets. In addition, because people with different age ranges, cultures and genders display and interpret emotion expression in different ways, an ideal multimodal dataset is expected to include abundant samples with precise attribute labels and other attributes such as age, gender and ethnicity, which can facilitate related research on age-invariant, gender-invariant and cultural-invariant emotion recognition using deep learning techniques. Therefore, it is essential to collect in-the-wild datasets with accurate annotations of emotion and demographic information, and plenty of subjects from the whole world for better generalization on the task of emotion recognition.

## 6 CONCLUSION

This paper comprehensively reviews and summarizes the definition of emotion models and the state-of-the-art of unimodal emotion recognition including facial expression recognition, speech emotion recognition and textual emotion recognition in dynamic data. In addition, this paper summarizes corresponding benchmark datasets, metrics and performances for clearly comprehending the development trend of research on the issue of emotion recognition. Ultimately, we present the latent research challenge and future direction to enrich the research in this field.

TABLE 3
Summary of the state-of-the-art in unimodality proposed in recent three years. F=face, S=speech, T=text.

| Modality | Method | Year | Datasets | Metrics | Performance |
|---|---|---|---|---|---|
| F | D. Deng et al. [64] | 2020 | OMG | CCC | Valence0.529 Arousal0.377 |
| | D. Kollias et al. [42] | 2020 | OMG | CCC | Valence0.535 Arousal0.365 |
| | J. Han et al. [2] | 2019 | RECOLA<br>OMG | CCC, F1 | CCC: Valence0.516 Arousal0.475(RECOLA)<br>F1: 43.9(OMG, 7 classes) |
| S | W. Lin et al. [76] | 2021 | MSP-Podcast<br>MSP-IMPROV<br>IEMOCAP | CCC | Valence0.2856 Arousal0.7027 Dominance0.6201(MSP-Podcast)<br>Valence0.354 Arousal0.563 Dominance0.422(MSP-IMPROV)<br>Valence0.365 Arousal0.629 Dominance0.469(IEMOCAP) |
| | A. Shukla et al. [1] | 2021 | CREMA-D<br>Ravdess<br>IEMOCAP<br>SEWA<br>RECOLA | F1, CCC | F1: 0.592(CREMA-D, 6 classes)<br>F1: 0.645(Ravdess, 8 classes)<br>F1: 0.642(IEMOCAP, 4 classes)<br>CCC: Valence0.38 Arousal0.383(SEWA)<br>CCC: Valence0.452 Arousal0.764(RECOLA) |
| | Mustaqeem et al. [73] | 2021 | IEMOCAP<br>RAVDESS<br>EMO-DB | uACC,ACC,uF1,F1 | uACC: 0.8, ACC:0.78, uF1: 0.78, F1:0.79(IEMOCAP, 4 classes)<br>uACC:0.8, ACC:0.8, uF1: 0.8, F1:0.81(RAVDESS, 8 classes)<br>uACC: 0.93, ACC: 0.95, uF1: 0.93, F1: 0.94(EMO-DB, 7 classes) |
| | R. Chatterjee et al. [74] | 2021 | RAVDESS<br>TESS | ACC | 0.9048(RAVDESS, 8 classes)<br>0.9579(TESS, 7 classes) |
| | S. Li et al. [75] | 2021 | IEMOCAP<br>EMO-DB<br>eNTERFACE'05<br>SAVEE | uACC, ACC | uACC: 0.7198, ACC: 0.8047(IEMOCAP, 4 classes)<br>uACC: 0.821, ACC: 0.833(EMO-DB, 7 classes)<br>uACC: 0.756, ACC: 0.758(eNTERFACE'05, 6 classes)<br>uACC: 0.5475, ACC: 0.565(SAVEE, 7 classes) |
| | J. Hsu et al. [136] | 2021 | NNIME | ACC, F1 | ACC: 0.6192, F1: 0.59(6 classes) |
| | S. Latif et al. [3] | 2020 | IEMOCAP<br>MSP-IMPROV | uACC,ACC | uACC:0.677, ACC:0.685(IEMOCAP, 4 classes)<br>uACC:0.602, ACC:0.625(MSP-IMPROV, 4 classes) |
| | S. Parthasarathy et al. [137] | 2020 | MSP-Podcast | CCC | Valence0.301, Arousal0.77, Dominance0.7 |
| | J. Han et al. [2] | 2019 | RECOLA<br>OMG | CCC, F1 | CCC: Valence0.434 Arousal0.644(RECOLA)<br>F1: 41.7(OMG, 7 classes) |
| | S. Zhang et al. [72] | 2019 | BAUM-1s | ACC | 0.5022(7 classes) |
| | S. Khorram et al. [138] | 2019 | RECOLA<br>SEWA | CCC | Valence0.475 Arousal0.814(RECOLA)<br>Valence0.432 Arousal0.433(SEWA) |
| T | W. Shen et al. [4] | 2021 | IEMOCAP<br>MELD<br>DailyDialog | F1 | 0.6594(IEMOCAP, 6 classes)<br>0.6241(MELD, 7 classes)<br>0.5493(DailyDialog, 7 classes) |
| | D. Hazarika et al. [90] | 2021 | IEMOCAP<br>DailyDialog<br>SEMAINE | F1, PCC | F1: 59.8(IEMOCAP, 6 classes)<br>F1: 48.0(DailyDialog, 7 classes)<br>PCC: Valence0.66 Arousal0.42 Power0.35 Expectancy-0.02(SEMAINE) |
| | H. Fei et al. [139] | 2020 | SemEval2018<br>Ren-CECps | ACC | 0.568(SemEval2018, 11 classes)<br>0.751(Ren-CECps, 8 classes) |
| | W. Jiao et al. [89] | 2020 | IEMOCAP<br>MELD | ACC, uF1, F1 | ACC: 0.635, F1: 0.635, uF1: 0.63(IEMOCAP, 6 classes)<br>ACC: 0.603, F1: 0.581, uF1: 0.386(MELD, 7 classes) |
| | C. T. Heaton et al. [140] | 2020 | IEMOCAP | uACC,ACC | uACC:0.794, ACC:0.809(4 classes) |
| | J. Deng et al. [88] | 2020 | RenCECps | F1, uF1, uACC | F1: 0.6076, uF1: 0.4831, uACC: 0.7651(9 classes) |
| | S. Zhu et al. [141] | 2019 | EMOBANK | PCC | Valence(0.372) Activation(0.233) Dominance(0.194) |

## ACKNOWLEDGMENTS

This work is partially supported by the National Key Research and Development Program of China under Grant No.2019YFB1405803, and the National Natural Science Foundation of China under Grants No. 61772125.

## REFERENCES

[1] A. Shukla, S. Petridis, and M. Pantic, "Does visual self-supervision improve learning of speech representations for emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[2] J. Han, Z. Zhang, Z. Ren, and B. W. Schuller, "Emobed: Strengthening monomodal emotion recognition via training with cross-modal emotion embeddings," *IEEE Transactions on Affective Computing*, 2019.

[3] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2020.

[4] W. Shen, J. Chen, X. Quan, and Z. Xie, "Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 13 789–13 797. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17625

[5] L. Zheng, A. Bl, and C. Jtab, "Decn: Dialogical emotion correction network for conversational emotion recognition," *Neurocomputing*, 2021.

[6] D. Zhang, W. Zhang, S. Li, Q. Zhu, and G. Zhou, "Modeling both intra- and inter-modal influence for real-time emotion detection in conversations," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 503–511. [Online]. Available: https://doi.org/10.1145/3394171.3413949

[7] S. Xing, S. Mai, and H. Hu, "Adapted dynamic memory network for emotion recognition in conversation," *IEEE Transactions on Affective Computing*, 2020.

[8] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6818–6825, 2019.

[9] D. Zhang, L. Wu, C. Sun, S. Li, Q. Zhu, and G. Zhou, "Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, Ed. ijcai.org, 2019, pp. 5415–5421. [Online]. Available: https://doi.org/10.24963/ijcai.2019/752

[10] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2021.

[11] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[12] K. Patel, D. Mehta, C. Mistry, R. Gupta, S. Tanwar, N. Kumar, and M. Alazab, "Facial sentiment analysis using ai techniques: State-of-the-art, taxonomies, and challenges," *IEEE Access*, vol. 8, pp. 90 495–90 519, 2020.

[13] A. Mba and B. Ko, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers - sciencedirect," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[14] N. Alswaidan and M. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, no. 16, 2020.

[15] Y. Jiang, W. Li, M. S. Hossain, M. Chen, and M. Al-Hammadi, "A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition," *Information Fusion*, vol. 53, 2019.

[16] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[17] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2017.

[18] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[19] D'Mello, Sidney, K., Kory, and Jacqueline, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys*, 2015.

[20] Ekman and Paul, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[21] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots," 2001.

[22] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[23] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[24] P. Viola and 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, 2001, pp. I–I.

[25] "Active appearance models." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2001.

[26] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012.

[27] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3444–3451.

[28] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[29] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[30] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1859–1866.

[31] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.

[32] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[33] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, 1987.

[34] A. B. Watson, "Image compression using the discrete cosine transform," *Mathematica Journal*, vol. 4, no. 7, pp. 81–88, 1994.

[35] S. Dabbaghchian, A. Aghagolzadeh, and M. S. Moin, "Feature extraction using discrete cosine transform for face recognition," in *International Symposium on Signal Processing & Its Applications*, 2007.

[36] Y. Zhang, F. Xiong, and G. L. Zhang, "A preprocessing algorithm for illumination invariant face recognition," *Journal of Image and Graphics*, 2008.

[37] P. Birch, B. Mitra, N. M. Bangalore, S. Rehman, R. Young, and C. Chatwin, "Approximate bandpass and frequency response models of the difference of gaussian filter," *Optics Communications*, vol. 283, no. 24, pp. 4942–4948, 2010.

[38] J. Short, J. Kittler, and K. Messer, "A comparison of photometric normalisation algorithms for face verification," *Proceedings of Automatic Face & Gesture Recognition*, 2004.

[39] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," *IEEE*, 2014.

[40] A. Yao, D. Cai, H. Ping, S. Wang, and Y. Chen, "Holonet: towards robust emotion recognition in the wild," in *Acm International Conference on Multimodal Interaction*, 2016.

[41] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," ser. ICMI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 553–560. [Online]. Available: https://doi.org/10.1145/3136755.3143009

[42] D. Kollias and S. P. Zafeiriou, "Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[43] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-gcn: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Transactions on Multimedia*, 2020.

[44] M. Dahmane, J. Alam, P.-L. St-Charles, M. Lalonde, K. Heffner, and S. Foucher, "A multimodal non-intrusive stress monitoring from the pleasure-arousal emotional dimensions," *IEEE Transactions on Affective Computing*, 2020.

[45] S. Peng, L. Zhang, Y. Ban, M. Fang, and S. Winkler, "A deep network for arousal-valence emotion prediction with acoustic-visual cues," 2018.

[46] D. Kollias and S. Zafeiriou, "A multi-component cnn-rnn approach for dimensional emotion recognition in-the-wild," 2018.

[47] D. Deng, Y. Zhou, J. Pi, and B. E. Shi, "Multimodal utterance-level affect analysis using visual, audio and text features," 2018.

[48] Z. Zheng, C. Cao, X. Chen, and G. Xu, "Multimodal emotion recognition for one-minute-gradual emotion challenge," 2018.

[49] A. Triantafyllopoulos, H. Sagha, F. Eyben, and B. Schuller, "audeering's approach to the one-minute-gradual emotion challenge," 2018.

[50] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[51] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *Eprint Arxiv*, 2014.

[52] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[53] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[54] Q. Li, D. Gkoumas, A. Sordoni, J. Nie, and M. Melucci, "Quantum-inspired neural network for conversational emotion recognition," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 13 270–13 278. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17567

[55] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and 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)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 873–883. [Online]. Available: https://aclanthology.org/P17-1081

[56] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[57] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[58] Y. Ma, Y. Hao, C. Min, J. Chen, L. Ping, and K. Andrej, "Audio-visual emotion fusion(avef):a deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2018.

[59] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and 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, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2122–2132. [Online]. Available: https://aclanthology.org/N18-1193

[60] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 2594–2604. [Online]. Available: https://aclanthology.org/D18-1280

[61] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[62] W. Kay, J. Carreira, K. Simonyan, B. Zhang, and A. Zisserman, "The kinetics human action video dataset," 2017.

[63] S. Zhao, Y. Ma, Y. Gu, J. Yang, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 1, pp. 303–311, 2020.

[64] D. Deng, Z. Chen, Y. Zhou, and B. E. Shi, "MIMAMO net: Integrating micro- and macro-motion for video emotion recognition," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 2020, pp. 2621–2628.

[65] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[66] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," pp. 292–301, 2018.

[67] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2017.

[68] X. Pan, G. Ying, G. Chen, H. Li, and W. Li, "A deep spatial and temporal aggregation framework for video-based facial expression recognition," *IEEE Access*, vol. 7, pp. 48 807–48 815, 2019.

[69] D. Feng and F. Ren, "Dynamic facial expression recognition based on two-stream-cnn with lbp-top," in *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, 2018.

[70] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," in *Advanced Techniques in Computing Sciences and Software Engineering*, K. Elleithy, Ed. Dordrecht: Springer Netherlands, 2010, pp. 279–282.

[71] J. Lin, C. Wu, and W. Wei, "Error weighted semi-coupled hidden markov model for audio-visual emotion recognition," *IEEE Trans. Multim.*, vol. 14, no. 1, pp. 142–156, 2012. [Online]. Available: https://doi.org/10.1109/TMM.2011.2171334

[72] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional lstm," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[73] Mustaqeem and S. Kwon, "Att-net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, no. 4, 2021.

[74] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, "Real-time speech emotion analysis for smart home assistants," *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, 2021.

[75] A. Sl, B. Xx, B. Wf, C. Bc, and B. Pf, "Spatiotemporal and frequential cascaded attention networks for speech emotion recognition," *Neurocomputing*, 2021.

[76] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, 2021.

[77] T. Mikolov, I. Sutskever, C. Kai, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013.

[78] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *Computer Science*, 2013.

[79] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing*, 2014.

[80] M. Peters, M. Neumann, M. Iyyer, M. Gardner, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018.

[81] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[82] J. Camacho-Collados and M. T. Pilehvar, "From word to sense embeddings: A survey on vector representations of meaning," *Journal of Artificial Intelligence Research*, 2018.

[83] K. Song, X. Tan, T. Qin, J. Lu, and T. Y. Liu, "Mass: Masked sequence to sequence pre-training for language generation," 2019.

[84] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," 2019.

[85] P. Xu, A. Madotto, C. S. Wu, J. H. Park, and P. Fung, "Emo2vec: Learning generalized emotion representation by multi-task training," 2018.

[86] B. Felbo, A. Mislove, A. Sgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[87] G. I. Winata, A. Madotto, Z. Lin, J. Shin, Y. Xu, P. Xu, and P. Fung, "Caire_hkust at semeval-2019 task 3: Hierarchical attention for dialogue emotion classification," 2019.

[88] J. Deng and F. Ren, "Multi-label emotion detection via emotion-specified feature extraction and emotion correlation learning," *IEEE Transactions on Affective Computing*, 2020.

[89] W. Jiao, M. Lyu, and I. King, "Real-time emotion recognition via attention gated hierarchical memory network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 8002–8009, 2020.

[90] D. Hazarika, S. Poria, R. Zimmermann, and R. Mihalcea, "Conversational transfer learning for emotion recognition," *Information Fusion*, vol. 65, pp. 1–12, 2021.

[91] D. Zhang, X. Ju, W. Zhang, J. Li, S. Li, Q. Zhu, and G. Zhou, "Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 14 338–14 346. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17686

[92] G. Xu, W. Li, and J. Liu, "A social emotion classification approach using multi-model fusion," *Future Generation Computer Systems*, vol. 102, 2019.

[93] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2852–2861. [Online]. Available: https://doi.org/10.1145/3394171.3413579

[94] M. Jaiswal and E. M. Provost, "Privacy enhanced multimodal neural representations for emotion recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 7985–7993, 2020.

[95] R. Li, Z. Wu, J. Jia, Y. Bu, and H. Meng, "Towards discriminative representation learning for speech emotion recognition," in *Twenty-Eighth International Joint Conference on Artificial Intelligence IJCAI-19*, 2019.

[96] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *arXiv*, 2017.

[97] P. Tzirakis, J. Chen, S. Zafeiriou, and B. Schuller, "End-to-end multimodal affect recognition in real-world environments," *Information Fusion*, vol. 68, pp. 46–53, 2021.

[98] T. Mittal, U. Bhattacharya., R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[99] Z. Lian, B. Liu, and J. Tao, "Ctnet: Conversational transformer network for emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 985–1000, 2021.

[100] A. Zadeh, S. Poria, P. P. Liang, E. Cambria, N. Mazumder, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," New Orleans, LA, United states, 2018, pp. 5634 – 5641, attention mechanisms;Benchmark datasets;Multi-views;Neural architectures;Sequential learning;Specific interaction;State of the art;.

[101] E. Mansouri-Benssassi and J. Ye, "Synch-graph: Multisensory emotion recognition through neural synchrony via graph convolutional networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 2, pp. 1351–1358, 2020.

[102] M. Wu, W. Su, L. Chen, W. Pedrycz, and K. Hirota, "Two-stage fuzzy fusion based-convolution neural network for dynamic emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[103] Y. Zhang, Z. R. Wang, and J. Du, "Deep fusion: An attention guided factorized bilinear pooling for audio-video emotion recognition," 2019.

[104] F. Shuang and C. Chen, "Fuzzy broad learning system: A novel neuro-fuzzy model for regression and classification," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–11, 2018.

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

[106] "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[107] B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Acm International Conference on Multimodal Interaction*, 2012.

[108] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and 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)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. [Online]. Available: https://www.aclweb.org/anthology/P18-1208

[109] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018.

[110] S. Poria, D. Hazarika, N. Majumder, G. Naik, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[111] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "Nnime: The nthu-ntua chinese interactive multimodal emotion corpus," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 292–298.

[112] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, 2021.

[113] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *International Conference on Data Engineering Workshops*, 2006.

[114] S. Haq, P. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. on Auditory-Visual Speech Processing (AVSP'08), Tangalooma, Australia*, Sept. 2008.

[115] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, p. 0034, 2012.

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

[117] Y. G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," *AAAI Press*, 2014.

[118] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[119] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[120] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "Baum-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2017.

[121] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2017.

[122] B. Xu, Y. Fu, Y. G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Transactions on Affective Computing*, vol. 9, no. 99, pp. 255–270, 2018.

[123] S. R. Livingstone, F. A. Russo, and N. Joseph, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. e0 196 391–, 2018.

[124] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *INTERSPEECH*, 2005.

[125] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[126] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: https://doi.org/10.5683/SP2/E8H2MF

[127] "Sentence emotion analysis and recognition based on emotion words using ren-cecps," *International Journal of Advanced Intelligence Paradigms*, vol. 2, no. 1, pp. 105–117, 2010.

[128] S. Buechel and U. Hahn, "Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis," in *EACL 2017*, 2017.

[129] Y. Li, S. Hui, X. Shen, W. Li, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," 2017.

[130] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[131] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://www.aclweb.org/anthology/L18-1252

[132] X. Ju, D. Zhang, J. Li, and G. Zhou, "Transformer-based label set generation for multi-modal multi-label emotion detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 512–520. [Online]. Available: https://doi.org/10.1145/3394171.3413577

[133] P. Barros, E. Barakova, and S. Wermter, "Adapting the interplay between personalized and generalized affect recognition based on an unsupervised neural framework," *IEEE Transactions on Affective Computing*, 2020.

[134] S. Zhou, J. Jia, Z. Wu, Z. Yang, Y. Wang, W. Chen, F. Meng, S. Huang, J. Shen, and X. Wang, "Inferring emotion from large-scale internet voice data: A semi-supervised curriculum

augmentation based deep learning approach," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 6039–6047. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/16753

[135] Y. Gu, X. Lyu, W. Sun, W. Li, S. Chen, X. Li, and I. Marsic, "Mutual correlation attentive factors in dyadic fusion networks for speech emotion recognition," ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 157–166. [Online]. Available: https://doi.org/10.1145/3343031.3351039

[136] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech emotion recognition considering nonverbal vocalization in affective conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1675–1686, 2021.

[137] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, 2020.

[138] S. Khorram, M. McInnis, and E. Mower Provost, "Jointly aligning and predicting continuous emotion annotations," *IEEE Transactions on Affective Computing*, 2019.

[139] H. Fei, Y. Zhang, Y. Ren, and D. Ji, "Latent emotion memory for multi-label emotion classification," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 5, pp. 7692–7699, 2020.

[140] C. T. Heaton and D. M. Schwartz, "Language models as emotional classifiers for textual conversation," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2918–2926. [Online]. Available: https://doi.org/10.1145/3394171.3413755

[141] S. Zhu, S. Li, and G. Zhou, "Adversarial attention modeling for multi-dimensional emotion regression," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 471–480. [Online]. Available: https://www.aclweb.org/anthology/P19-1045

[142] P. Ekman, "Constants across cultures in the face and emotion." *J Pers Soc Psychol*, vol. 17, no. 2, pp. 124–129, 1971.

[143] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sincadk, and P. Dario, "Emotion modelling for social robotics applications: A review," *Journal of Bionic Engineering*, no. 2, pp. 185–203, 2018.

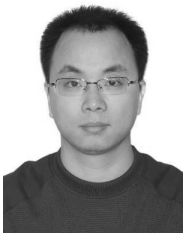
[144] Stephan, Hamann, , , Turhan, and Canli, "Individual differences in emotion processing," *Current Opinion in Neurobiology*, 2004.

[145] U. Hess, C. Blaison, and K. Kafetsios, "Judging facial emotion expressions in context: The influence of culture and self-construal orientation," *Journal of Nonverbal Behavior*, vol. 40, no. 1, pp. 55–64, 2016.

[146] J. Gideon, M. McInnis, and E. Mower Provost, "Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (addog)," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[147] H. Luo and J. Han, "Nonnegative matrix factorization based transfer subspace learning for cross-corpus speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2047–2060, 2020.

**Tao Zhang** received the master's degree from Software College, Northeastern University, Shenyang, China, in 2020, where he is currently pursuing the Ph.D degree. He has authored or coauthored 5 publications in journals and conferences. He holds 4 Chinese patents. His current research interests include emotion reognition, biometric recognition, blind source separation and lifelong learning.



**Zhenhua Tan** received the B.S., M.S., and Ph.D. degrees from Northeastern University, Shenyang, China, in 2003, 2006, and 2009, respectively, all in computer science. He is a professor and a PhD Supervisor with Software College now. He is also of the High-level talent in Shenyang, first batch of Industrial Information Security Expert in Liaoning province. He has authored or coauthored over 50 publications in journals, conferences and book chapters. He holds 3 US international patents (all granted). His current research interests include cyber security, cyberspace data analysis, verification and identification technologies based on AI. He was funded by the National Natural Science Foundation of China, Ministry of Education of China, Natural Science Foundation of Liaoning, and Industries, with more than 10 projects.