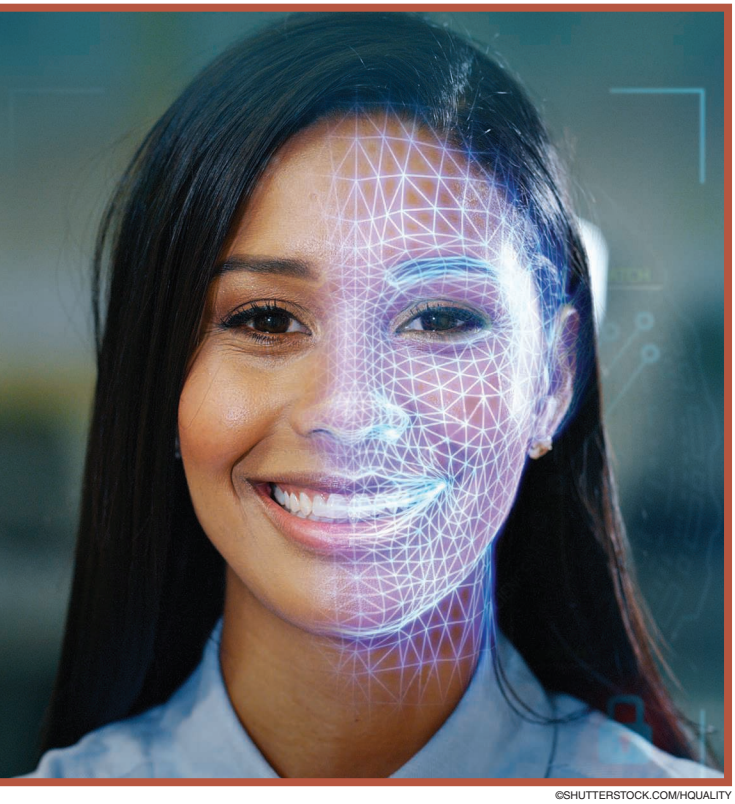
Sicheng Zhao, Guoli Jia, Jufeng Yang, Guiguang Ding, and Kurt Keutzer

Emotion Recognition From Multiple Modalities

Fundamentals and methodologies



©SHUTTERSTOCK.COM/HQUALITY

Digital Object Identifier 10.1109/MSP.2021.3106895
Date of current version: 27 October 2021

Humans are emotional creatures. Multiple modalities are often involved when we express emotions, whether we do so explicitly (such as through facial expression and speech) or implicitly (e.g., via text or images). Enabling machines to have emotional intelligence, i.e., recognizing, interpreting, processing, and simulating emotions, is becoming increasingly important. In this tutorial, we discuss several key aspects of multimodal emotion recognition (MER).

We begin with a brief introduction on widely used emotion representation models and affective modalities. We then summarize existing emotion annotation strategies and corresponding computational tasks, followed by a description of the main challenges in MER. Furthermore, we present some representative approaches on representation learning of each affective modality, feature fusion of different affective modalities, and classifier optimization as well as domain adaptation for MER. Finally, we outline several real-world applications and discuss some future directions.

Introduction

Emotion is present everywhere in human daily life and can influence or even determine our judgment and decision making [1]. For example, in marketing, a widely advertised brand can generate a mental representation of a product in consumers' minds and influence their preferences and actions; inducing sadness and disgust during a shopping trip would, respectively, increase and decrease consumers' willingness to pay [31]. Drivers experiencing strong emotions, such as sadness, anger, agitation, and even happiness, are much more likely to be involved in an accident [32]. In education—especially current online classes during the COVID-19 pandemic period—students' emotional experiences and interactions with teachers have a big impact on their learning ability, interest, engagement, and even career choices [33].

The importance of emotions in artificial intelligence was recognized decades ago. Minsky, a Turing Award winner in 1970, once claimed, "The question is not whether intelligent machines can have any emotions, but whether machines can be

intelligent without emotions" [2]. Enabling machines to have emotional intelligence, i.e., recognizing, interpreting, processing, and simulating emotions, has recently become increasingly important, with wide potential applications involving human–computer interaction [3].

On the one hand, emotionally intelligent machines can provide more harmonious and personal services for human beings, especially the elderly, those with disabilities, and children. For example, companion robots that can work with emotions can better meet the psychological and emotional needs of the elderly and help them stay comfortable.

On the other hand, by recognizing humans' emotions automatically and in real time, intelligent machines can better identify humans' abnormal behaviors, send reminders to their relatives and friends, and prevent extreme behaviors to themselves and even to the rest of society. For example, an emotion-monitoring system for driving can automatically play some soothing music to relax angry individuals who might be dissatisfied with a traffic jam and can remind them to focus on driving safely.

The first step for intelligent machines to express human-like emotions is to recognize and understand humans' emotions, typically through two groups of affective modalities: explicit affective cues and implicit affective stimuli. Explicit affective cues correspond to specific physical and psychological changes in humans that can be directly observed and recorded, such as facial expressions, eye movement, speech, actions, and physiological signals. These can be either easily suppressed and masked or difficult and impractical to capture.

Meanwhile, the popularity of mobile devices and social networks enables humans to habitually share their experiences and express their opinions online using text, images, audio, and video. Implicit affective stimuli correspond to these commonly used digital media, the analysis of which provides an implicit way to infer humans' emotions [4].

Regardless of whether emotions are expressed explicitly or implicitly, there are generally multiple modalities that can contribute to the emotion recognition task, as shown in Figure 1. As compared to unimodal emotion recognition, MER has several advantages. The first is data complementarity. Cues from different modalities can augment or complement each other. For example, if we see a social media post from a good friend saying, "What great weather!" it is highly probable that our friend is expressing a positive emotion, but, if there is also an auxiliary image of a storm, we can infer that the text is actually sarcastic and that a negative emotion is intended to be expressed.

The second is model robustness. Due to the influence of many normally occurring factors in data collection, such as sensor device failure, some data modalities might be unavailable, which is especially prevalent in the wild. For example, in the CALLAS data set containing speech, facial expression, and gesture modalities, the gesture stream is missing for some momentarily motionless users [5]. In such cases, the learned MER model can still work with the help of other available modalities.

The final advantage is performance superiority. Joint consideration of the complementary information of different modalities can result in better recognition performance. A meta-analysis indicates that, as compared to the best unimodal counterparts, MER achieves 9.83% performance improvement on average [6].

In this article, we give a comprehensive tutorial on different aspects of MER, including psychological models, affective modalities, data collections and emotion annotations, computational tasks, challenges, computational methodologies, applications, and future directions. There have been several reviews/surveys on MER-related topics [4], [6]–[9]. In particular, [7] and [9] cover different aspects of general multimodal machine learning with few efforts on emotion recognition, [6] focuses on the quantitative review and meta-analysis of existing MER systems, and [4] and [8] are survey-style MER articles with a technical emphasis on multimodal fusion. However, this tutorial-style article aims to give a quick and comprehensive MER introduction that is also suitable for nonspecialists.

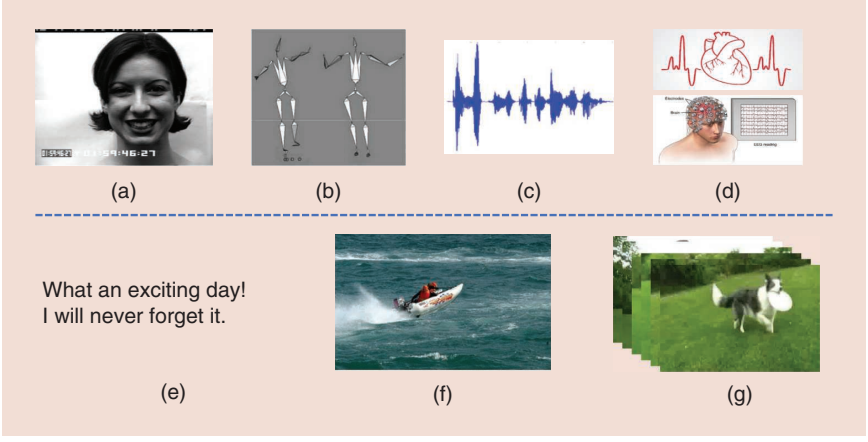


FIGURE 1. The multiple modalities for emotion recognition. Explicit affective cues include (a) facial expression, (b) action and gait, (c) speech, and (d) physiological signals. Implicit affective stimuli include (e) text, (f) image, and (g) video.

Psychological models

In psychology, categorical emotion states (CES) and dimensional emotion space (DES) are two representative types of models to measure emotion [10]. CES models define emotions as being in a few basic categories, such as binary sentiments (positive and negative, sometimes including neutral), Ekman's six basic emotions [happiness and surprise (positive) as well as anger, disgust, fear, and sadness (negative)], Mikels's eight emotions [amusement, awe, contentment, and excitement (positive) as well as anger, disgust, fear, and sadness (negative)], Plutchik's emotion wheel (eight basic emotion categories, each

with three intensities), and Parrott's tree hierarchical grouping (primary, secondary, and tertiary categories). The development of psychological theories motivates CES to be increasingly diverse and fine-grained. DES models employ continuous 2D, 3D, or higher-dimensional Cartesian spaces to represent emotions; the most widely used DES model is valence–arousal–dominance (VAD), which represent the pleasantness, intensity, and control degree of emotions, respectively.

CES models agree better with humans' intuition, but no consensus has been reached by psychologists on how many discrete emotion categories should be included. Furthermore, emotion is complex and subtle, which cannot be well reflected by limited discrete categories. DES models can theoretically measure all emotions as different coordinate points in the continuous Cartesian space, but the absolute continuous values are beyond users' understanding. These two types of definitions of emotions are related, with a possible transformation from a CES to DES. For example, anger relates to negative valence, high arousal, and high dominance.

Besides emotion, there are several other widely used concepts in affective computing, such as mood, affect, and sentiment. Emotions can be expected, induced, or perceived. We do not aim to distinguish them in this article. Please refer to [11] for more details on the differences or correlations among these concepts.

Affective modalities

In the area of MER, multiple modalities are employed to recognize and predict human emotions. The affective modalities in MER can be roughly divided into two groups based on whether emotions are recognized from humans' physical body changes or external digital media: explicit affective cues and implicit affective stimuli.

The former group includes facial expression, eye movement, speech, action, gait, and electroencephalography (EEG), all of which can be directly observed, recorded, or collected from an individual. Meanwhile, the latter group comprises commonly used digital media types, such as text, audio, images, and video. We use these data types to store information and knowledge as well as transfer them among digital devices. In this way, emotions may be implicitly involved and evoked. Although the efficacy of one specific modality as a reliable channel to express emotions cannot be guaranteed, jointly considering multiple modalities would significantly improve the reliability and robustness [12].

Explicit affective cues

A facial expression is an isolated motion of one or more human face regions/units or a combination of such motions. It is commonly agreed that facial expressions can carry informative affective cues, and they are recognized as one of the most natural and powerful signals to convey the emotional states and intentions of humans [12]. Facial expression is also a form of nonverbal communication conveying social information among humans.

We can deduce how an individual is feeling by observing his or her eye movement [34]. The eyes are often viewed as important cues of emotions. For example, if a person is nervous or lying, the blinking rate of his or her eyes may become slower than normal [34]. Eye movement signals can be easily collected via an eye-tracker system and have been widely used in human–computer interaction research.

Speech is a significant vocal modality to carry emotions [13], [14]. Speakers may express their intentions, like asking or declaring, by using various intonations, degrees of loudness, and tempo. Specifically, emotions can be revealed when people talk with each other or just mutter to themselves.

As an important part of human body language, action also conveys massive information about emotion. For instance, an air punch is an act of thrusting one's clenched fist up into the air, typically as a gesture of triumph or elation.

Similar to action, emotions can be perceived from a person's gait, i.e., his or her walking style. The psychology literature has proven that participants can identify the emotions of a subject by observing his or her posture, including long strides, collapsed upper body, and so on [35]. Body movement (e.g., walking speed) also plays an important role in the perception of different emotions. High-arousal emotions, such as anger and excitement, are more associated with rapid movements than low-arousal emotions, such as sadness and contentment.

Last but not least, EEG, as one representative psychological signal, is another important method for recording the electrical and emotional activity of the brain [15]. Compared to the other aforementioned explicit cues, the collection of EEG signals is typically more difficult and unnatural, regardless of whether electrodes are placed noninvasively along the scalp or invasively using electrocorticography.

Implicit affective stimuli

Text is a form used to record the natural language of human beings, which can implicitly carry informative emotions [16], [17]. It has different levels of linguistic components, including words, sentences, paragraphs, and articles, which are well studied; many off-the-shelf algorithms have been developed to segment text into small pieces. Then, the affective attribute of each linguistic piece is recognized with the help of a publicly available dictionary like SentiWordNet, and the emotion evoked by the text can be deduced.

A digital audio signal is a representation of sound, typically stored and transferred using a series of binary numbers [12]. Audio signals may be synthesized directly or originate at a transducer, such as a microphone or musical instrument. Unlike speech, which mainly focuses on human vocal information and the content of which may be translated into natural language, audio is more general, including any sound, like music or birdsong.

An image is a distribution of colored dots over space [36]. The phrase "a picture is worth a thousand words" is well known. It has been demonstrated in psychology that emotions can be evoked in humans by images [18]. The

explosive growth of images shared online and powerful descriptive ability of scenes have enabled images to become crucial affective stimuli, which has attracted extensive research efforts [10].

Video naturally contains multiple modalities at the same time, such as visual, audio, and textual information [19]. That means temporal, spatial, and multichannel representations can be learned and utilized to recognize the emotions in videos.

Data collections and emotion annotations

Two steps are usually involved in constructing an MER data set: data collection and emotion annotation. The collected data can be roughly divided into two categories: selecting from existing data and new recording in specific environments.

On the one hand, some data are selected from movies, reviews, videos, and TV shows in online social networks, such as YouTube and WeiBo. For example, the review videos in ICT-MMMO and MOUD are collected from YouTube; audiovisual clips are extracted from TV series in MELD; online reviews from the food and restaurant categories are crawled in Yelp; and video blogs, typically with one speaker looking at the camera from YouTube, are collected in CMU-MOSI to capture the speakers' information. Some collected data provide a transcription of speech either manually (e.g., CMU-MOSI and CH-SMIS) or automatically (such as ICT-MMMO and MELD).

On the other hand, some data are newly recorded with different sensors in specifically designed environments. For example, participants' physiological signals and frontal facial changes induced by music videos are recorded in DEAP.

There are different kinds of emotion annotation strategies. Some data sets have target emotions and do not need to be annotated. For example, in EMODB, each sentence performed by actors corresponds to a target emotion. For some data sets, the emotion annotations are obtained automatically. For example, in Multi-ZOL, the integer sentiment score for each review, ranging from 1 to 10, is regarded as the sentiment label.

Several workers are employed to annotate the emotions, such as VideoEmotion-8. The data sets with recorded data are usually annotated by participants' self-reporting, such as MAHNOB-HCI. In addition, the emotion labels are typically obtained by major voting.

For DES models, "FeelTrace" and "SAM" are often used for annotation. The former is based on the activation-evaluation space, which allows observers to track the emotion content of a stimulus as they perceive it over time. The latter is a tool that accomplishes emotion rating based on different Likert scales. Some commonly used data sets are summarized in Table 1.

Computational tasks

Given multimodal affective signals, we can conduct different MER tasks, including classification, regression, detection, and retrieval. In this section, we briefly introduce what these tasks do.

Emotion classification

In the emotion classification task, we assume that one instance can belong to only one or a fixed number of emotion categories, and the goal is to discover class boundaries or distributions in the data space [16]. Current works mainly focus on the manual design of multimodal features and classifiers or employing deep neural networks in an end-to-end manner.

As defined as a single-label learning problem, MER assigns a single dominant emotion label to each sample. However, the emotion may be a mixture of all components from various regions or sequences rather than a single representative emotion. Meanwhile, different people may have varying emotional reactions to the same stimulus, which is caused by a variety of elements, like personality.

Thus, multilabel learning (MLL) has been utilized to study the problem where one instance is associated with multiple emotion labels. Recently, to address the problem that MLL does not fit some real applications well where the overall distribution of different labels' importance matters, label-distribution learning is proposed to cover a certain number of labels, representing the degree to which each emotion label describes the instance [20].

Emotion regression

Emotion regression aims to learn a mapping function that can effectively associate one instance with continuous emotion values in a Cartesian space. The most common regression algorithms for MER aim to assign the average dimension values to the instance. To deal with the inherent subjectivity characteristic of emotions, researchers propose predicting the continuous probability distribution of emotions, which are represented in dimensional VA space. Specifically, VA emotion labels can be represented by a Gaussian mixture model (GMM), and then the emotion distribution prediction can be formalized as a parameter learning problem [21].

Emotion detection

As the raw data do not ensure carrying emotions, or only part of the data can evoke emotional reactions, emotion detection aims to find out which kind of emotion lies where in the source data. For example, a restaurant review on Yelp might read, "This location is conveniently located across the street from where I work—being walkable is a huge plus for me! Foodwise, it's the same as almost every location I've visited, so there's nothing much to say there. I do have to say that the customer service is hit or miss." Meanwhile, the overall rating score is three stars out of five. This review contains different emotions and attitudes: positive in the first sentence, neutral in the second sentence, and negative in the last sentence. As such, it is crucial for the system to detect which sentence corresponds to each emotion. Another example is affective region detection in images [22].

Emotion retrieval

How to search affective content based on human perception is another meaningful task. The existing framework first detects local interest patches or sequences in the query and candidate data sources. Then, it discovers all matched pairs by determining whether the distance between two patches or sequences is less than a given fixed threshold. The similarity score between the query and each candidate is calculated as the quantity of matched components, followed by ranking the candidates of this query accordingly. While an affective retrieval system is useful

Table 1. A brief summary of released data sets for MER.

Data Set	Modalities	Samples	Data Sources	Emotion Labels	Website
IEMOCAP	Face, speech, t-text, and video	10,039 turns	Recording	ang, sad, hap, dis, fea, sur, fru, exc, and neu VAD on 5-point ratings	https://sail.usc.edu/iemocap
YouTube	Face, eye, speech, t-text, and video	47 videos	YouTube	pos, neg, and neu	http://multicomp.cs.cmu.edu/rsources/youtube-dataset-2
MOUD	Face, speech, t-text, and video	412 utterances	YouTube	pos, and neg	http://web.eecs.umich.edu/~mihalcea/downloads.html#MOUD
ICT-MMMO	Face, eye, speech, t-text, and video	370 segments	Youtube and ExpoTV	pos and neg	http://multicomp.cs.cmu.edu/resources/ict-mmmo-dataset
News Rover	Face, speech, t-text, and video	929 videos	News	pos, neg, and neu	https://www.ee.columbia.edu/ln/dvmm/newsrover/sentimentdataset
CMU-MOSI	Face, eye, speech, t-text, and video	2,199 clips	YouTube	–3 to 3 sentiment score	http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset
CMU-MOSEI	Face, eye, speech, t-text, and video	23,453 sentences	YouTube	hap, sad, ang, fea, dis, and sur –3 to 3 sentiment score	http://multicomp.cs.cmu.edu/resources/cmu-mosei-dataset/
MELD	Face, speech, t-text, and video	13,708 utterances	TV series *Friends*	hap, sad, ang, fea, dis, sur, neu, and non-neu pos, neg, and neu	https://affective-meld.github.io
CH-SIMS	Face, eye, speech, t-text, and video	2,281 segments	Movies and TV series Variety shows	–1 to 1 sentiment score	https://github.com/thuiar/MMSA
eNTERFACE'05	Face, speech, and video	1,166 sequences	Recording	ang, fea, hap, sad, and sur	http://www.enterface.net/enterface05
SEMAINE	Face, speech, t-text, and video	959 conversations	Recording	val, act, pow, exp, int; bas-em, eps, ipa, and vad	https://semaine-db.eu/
EMDB	Video, SCL, and HR	52 clips	Films	ero, hor, neg, pos, sce, and obm VAD on 9-point ratings	EMDB@psi.uminho.pt
DEAP	Face, EEG, GSR, RA, and ST ECG, BVP, EMG, and EOG	1,280 samples	Recording	VAD-L on 9 point ratings F on 5-point ratings	http://www.eecs.qmul.ac.uk/mmv/datasets/deap/
MAHNOB-HCI	Face, eye, audio, and EEG ECG, GSR, ST, and RA	532 samples	Recording	sad, joy, dis, neu, hap, amu, ang, fea, sur, and anx VAD-P on 9-point ratings	https://mahnob-db.eu/hci-tagging
Multi-ZOL	Image and text	28,469 aspect-review pairs	ZOL	0–10 sentiment score	https://github.com/xunan0812/MIMN
Yelp	Image and text	244,569 images and 44,305 reviews	Yelp	Sentiment score on 5-point ratings	https://github.com/PreferredAI/vista-net
Tourism	Image and text	1,796 weibos	WeiBo	pos, neg, and neu	https://github.com/wlj961012/Multi-Modal-Event-awareNetwork-for-SentimentAnalysis-in-Tourism
LIRIS-ACCEDE	Video (audio and image)	9,800 clips	Movies	Rank along valence	https://liris-accede.ec-lyon.fr
VideoEmotion-8	Video (audio and image)	1,101 videos	YouTube and Flickr	ang, ant, dis, fea, joy, sad, sur, and tru	http://www.yugangjiang.info/research/VideoEmotions/index.html
Ekman-6	Video (audio and image)	1,637 videos	YouTube and Flickr	ang, dis, fea, joy, sad, and sur	https://github.com/kittenish/Frame-Transformer-Network

Modalities: BVP: blood volume pressure; ECG: electrocardiogram; EMG: electromyogram; EOG: electro-oculogram; GSR: galvanic skin response; HR: heart rate; PPS: peripheral physiological signal; RA: respiration amplitude; SCL: skin conductance level; ST: skin temperature; t-text: transcript text.
Emotion labels: amu: amusement; ang: angry; ant: anticipation; anx: anxiety; dis: disgust; ero: erotic; exc: excited; F: familiarity; fea: fear; fru: frustration; hap: happiness; hor: horror; L: liking; neg: negative; neu: neutral; obm: object manipulation; P: predictability; pos: positive; sad: sadness; sce: scenery; sur: surprise; tru: trust; act: activation; bas-em: basic-emotions; eps: epistemic-states; exp: expectation; ipa: interaction-process-analysis; int: intensity; pow: power; val: valence; vad: validity.

for obtaining online content with the desired emotions from a massive repository [10], again, the abstract and subjective characteristics make the task challenging and difficult to evaluate.

Challenges

As stated in the "Introduction" section, MER has several advantages as compared to unimodel emotion recognition, but it also faces more challenges.

Affective gap

The affective gap, which measures the inconsistency between extracted features and perceived high-level emotions, is one main challenge for MER. The affective gap is even more challenging than the semantic gap in objective multimedia analysis. Even if the semantic gap is bridged, there might still exist an affective gap.

For example, a blooming and a faded rose both contain a rose, but they can evoke different emotions. For the same sentence, different voice intonations may correspond to totally different emotions. Extracting discriminative high-level features, especially those related to emotions, can help to bridge the affective gap. The main difficulty lies in how to evaluate whether the extracted features are related to emotions.

Perception subjectivity

Due to many personal, contextual, and psychological factors, such as the cultural background, personality, and social context, different people might have varying emotional responses to the same stimuli [10]. Even if the emotion is the same, their physical and psychological changes can also be quite divergent.

For example, all of the 36 videos in the ASCERTAIN data set for MER are labeled with at least four out of seven different valence and arousal scales by 58 subjects [15]. This clearly indicates that some subjects have the opposite emotional reactions to the same stimuli. Take a short video with a storm and thunder, for instance: some people may feel awe because they have never seen such extreme weather, others may experience fear because of the loud thunder noise, some may be excited to capture such rare scenes, still others may feel sad because they have to cancel their travel plans, and so on.

Even for the same emotion (e.g., excitement), there are different reactions, such as facial expression, gait, action, and speech. For the subjectivity challenge, one direct solution is to learn personalized MER models for each subject. From the perspective of stimuli, we can also predict the emotion distribution when a certain number of subjects are involved. Besides the content of the stimuli and direct physical and psychological changes, jointly modeling the personal, contextual, and psychological factors mentioned earlier would also contribute to the MER task.

Data incompleteness

Because of the presence of many inevitable factors in data collection, such as sensor device failure, the information in specific modalities might be corrupted, which results in missing or incomplete data. Data incompleteness is a common phenomenon in real-world MER tasks.

For example, for explicit affective cues, an EEG headset might record contaminated signals or even fail to capture any signal; at night, cameras cannot capture clear facial expressions. For implicit affective stimuli, one user might post a tweet containing only an image (without text); for some videos, the audio channel does not change much. In such cases, the simplest feature fusion method, i.e., early fusion, does not work because we cannot extract any features given no captured signal. Designing effective fusion methods that can deal with data incompleteness is a widely employed strategy.

Cross-modality inconsistency

Different modalities of the same sample may conflict with each other and, thus, express varying emotions. For example, facial expressions and speech can be easily suppressed or masked to avoid being detected, but EEG signals that are controlled by the central nervous system can reflect humans' unconscious body changes. When people post tweets on social media, it is very common that the images are not semantically correlated to the text. In such cases, an effective MER method is expected to automatically evaluate which modalities are more reliable, such as by assigning a weight to each one.

Cross-modality imbalance

In some MER applications, different modalities may contribute unequally to the evoked emotion. For example, online news plays an important role in our daily lives, and, in addition to understanding the preferences of readers, predicting their emotional reactions is of great value in various applications, such as personalized advertising. However, a piece of online news usually includes imbalanced texts and images; i.e., an article may be very long, with lots of detailed information, while only one or two illustrations are inserted into the news. Potentially more problematic, the editor of the news may select a neutral image for an article with an obvious sentiment.

Label noise and absence

Existing MER methods, especially the ones based on deep learning, require large-scale labeled data for training. However, in real-world applications, labeling emotions in the ground-truth generation is not only prohibitively expensive and time-consuming but also highly inconsistent, which results in a large amount of data but with few or even no emotion labels. With the increasingly diverse and fine-grained emotion requirement, we might have enough training data for some emotion categories but not for others. One alternate solution to manual annotation is to leverage the tags or keywords of social tweets as emotion labels, but such labels are incomplete and noisy. As such, designing effective algorithms for unsupervised/weakly supervised learning and few-/zero-shot learning can provide potential solutions.

Meanwhile, we might have sufficient labeled affective data in one domain, such as synthetic facial expression and speech. The problem turns to how to effectively transfer the trained MER model on the labeled source domain to another unlabeled target domain. The presence of a domain shift causes significant performance decay when a direct transfer is used [23]. Multimodal domain adaptation and domain generalization can help to mitigate such domain gaps. Practical settings, such as multiple source domains, should also be considered.

Computational methodologies

Generally, there are three components in an MER framework with sufficient labeled training data in the target domain: representation learning, feature fusion, and classifier optimization, as shown in Figure 2. In this section, we introduce these components. Further, we describe domain adaptation when there is no labeled training data in the target domain and sufficient labeled data are available in another related source domain.

Representation learning of each affective modality

To represent text in a form that can be understood by computers, the following aspects are required: first, representing the symbolic words as real numbers for the next computation; second, modeling the semantic relationships; and, finally, obtaining a unified representation for the whole text [16]. In the beginning, words are represented by one-hot vectors with the length of the vocabulary size, where, for the tth word in the vocabulary w_t, only the position t is one, and the other positions are zero. As the scale of the data increases, the dimension of this one-hot vector increases dramatically.

Later, researchers used language models to train word vectors by predicting context, obtaining word vectors with vectors of a fixed dimension. Popular word vector representation models include word2vec, GLOVE, BERT, and XLNet, among others.

The text feature extraction methods have developed from simple to complex ones as well. Text features can be obtained by simply averaging word vectors. A recurrent neural network (RNN) is used to model the sequential relations of words in the text. A convolutional neural network (CNN), which has been widely employed in the computer vision community, is also used to extract the contextual relations between words.

To date, plenty of methods have been developed to design representative features for emotion stimuli in audios [13], [14]. It has been found that audio features, such as pitch, log energy, zero-crossing rate, spectral features, voice quality, and jitter, are useful in emotion recognition. The ComParE acoustic feature set has been commonly used as the baseline set for the ongoing Computation Paralinguistics Challenge series since 2013. However, because of possible high similarities in certain emotions, a single type of audio feature is not discriminative enough to classify emotions.

To solve this problem, some approaches propose combining different types of features. Recently, with the development of deep learning, CNNs are shown to achieve state-of-the-art

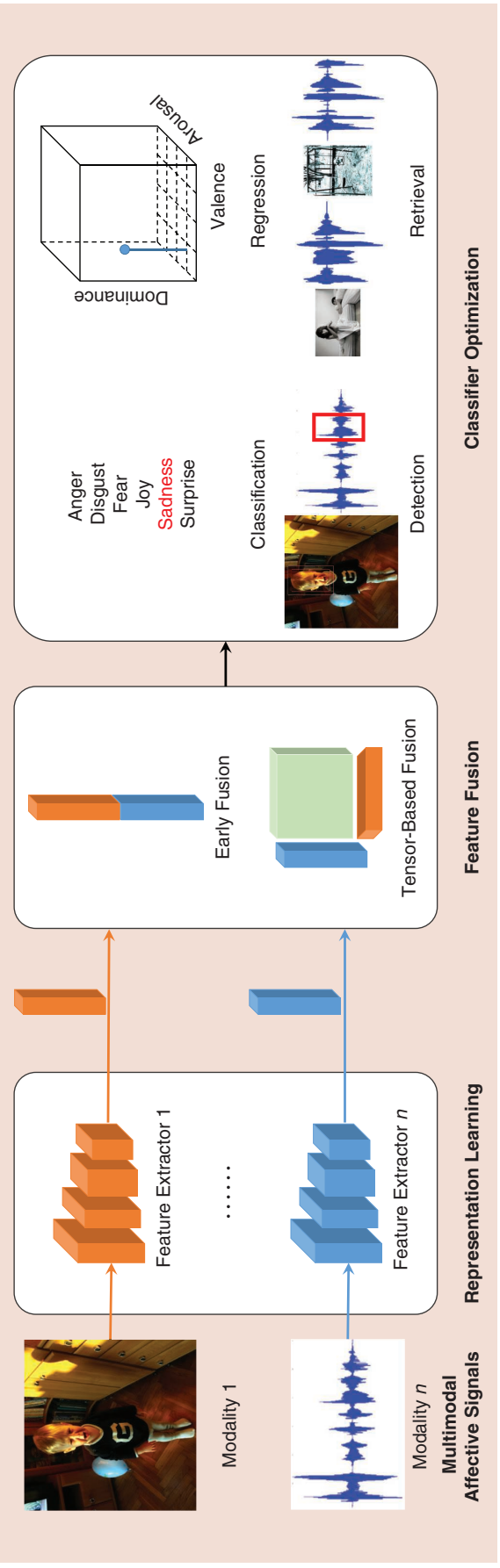


FIGURE 2. A widely used MER framework, which consists of three components: representation learning to extract feature representations, feature fusion to combine features from different modalities, and classifier optimization to learn specific task models (e.g., classification, regression, detection, and retrieval); n is the number of different modalities.

performance on large-scale tasks in many domains dealing with natural data, and audio emotion recognition is, of course, also included. Audio is typically transferred into a graphical representation, such as a spectrogram, to be fed into a CNN. Since the CNN uses shared weight filters and pooling to give the model better spectral and temporal invariant properties, it typically yields better generalized and more robust models for emotion recognition.

Researchers have designed informative representations for emotional stimuli in images. In general, images can be divided into two types, nonrestrictive and facial expression. For the former, e.g., natural images, various handcrafted features, including color, texture, shape, composition, and so on, are developed to represent image emotion in the early years [10]. These low-level features are developed with inspiration from psychology and art theory.

Later, midlevel features based on the visual concepts are presented to bridge the gap between the pixels in images and emotion labels. The most representative engine is SentiBank, which is composed of 1,200 adjective–noun pairs and shows remarkable and robust recognition performance among all of the hand-engineering features. In the era of deep learning, a CNN is regarded as a strong feature extractor in an end-to-end manner. Specifically, to integrate various representations of different levels, features are extracted from multiple layers of the CNN.

Meanwhile, an attention mechanism is employed to learn better emotional representations of specific local affective regions [22]. For the facial expression images, firstly, the human face is detected and aligned, and then the face landmarks are encoded for the recognition task. Note that, for those nonrestrictive images that contain human faces by chance, facial expression can be treated as an important midlevel cue.

Earlier, we mentioned how to identify emotions in isolated modalities. Here, we first focus on perceiving emotions from successive frames. Then, we introduce how to build joint representation for videos. Compared to a single image, a video contains a series of images with temporal information [19].

To build representations of videos, a wide range of methods has been proposed. Early methods mainly utilize handcrafted local representations in this field, which include color, motion, and the shot cut rate. With the advent of deep learning, recent methods extract discriminative representations by adopting a 3D CNN that captures the temporal information encoded in multiple adjacent frames. After extracting modality-specific features in videos, integrating different types of features could obtain more promising results and improve the performance.

To perceive emotions, there are mainly two aspects of ways to learn the representations of gait [24]. For one thing, we can explicitly model the posture and movement information that is related to the emotions. To do this, we first extract the skeletal structure of a person and then represent each joint of the human body using the 3D coordinate system. After getting these coordinates, the angles, distance, or area among different joints (posture information), velocity/acceleration (movement information), their covariance descriptors, and so on can be easily extracted.

For another thing, high-level emotional representations can be modeled from gait by long short-term memory (LSTM), deep CNNs, or graph convolutional networks. Some methods extract optical flow from gait videos and then extract sequence representations using these networks. Others learn skeletal structures of the gait and then feed them into multiple networks to extract discriminate representations.

Since various types of information about emotions, such as the frequency band, electrodeposition, and temporal data, can be explored from the brain's response to emotional stimuli, EEG signals are widely used in emotion analysis [15]. To extract discriminative features for EEG emotion recognition, differential entropy features from the frequency band or electrodeposition relationship are very popular in previous works.

In addition to handcrafted features, we can also directly apply end-to-end deep learning neural networks, such as CNNs and RNNs, on raw EEG signals to obtain powerful deep features [25]. Inspired by the learning pattern of humans, spatialwise attention mechanisms are successfully applied to extract more discriminative spatial information. Furthermore, considering that EEG signals contain multiple channels, a channelwise attention mechanism can also be integrated into a CNN to exploit the interchannel relationship among feature maps.

Feature fusion of different affective modalities

Feature fusion, as one key research topic in MER, aims to integrate the representations from multiple modalities to predict either a specific category or continuous value of emotions. Generally, there are two strategies: model-free and model-based fusion [7], [9].

Model-free fusion that is not directly dependent on specific learning algorithms has been widely used for decades. We can divide it into early fusion, late fusion, and hybrid fusion [5]. All of these fusion methods can be extended from existing unimodal emotion recognition classifiers.

Early fusion, also named *feature-level fusion*, directly concatenates the feature representations from different modalities as a single representation. It is the most intuitive method for fusing multiple representations by exploiting the interactions between various modalities at an early stage and only requires training a single model. However, since the representations from the modalities might significantly differ, we have to consider the time synchronization problem to transform these representations into the same format before fusion. When one or more modalities are missing, such early fusion would fail.

Late fusion, also named *decision-level fusion*, instead integrates the prediction results from each single modality. Some popular mechanisms include averaging, voting, and signal variance. The advantages of late fusion include 1) flexibility and superiority (the optimal classifiers can be selected for different modalities) and 2) robustness (when some modalities are missing, late fusion can still work). However, the correlations between different modalities before the decision are ignored.

Hybrid fusion combines early and late fusion to exploit their advantages in a unified framework but with higher computational cost.

Model-based fusion that explicitly performs fusion during the construction of learning models has received more attention [7], [9], as shown in Figure 3, since it is based on some simple techniques that are not specifically designed for multimodal data. For shallow models, kernel- and graph-based fusion are two representative methods; for recent popular deep models, neural network-, attention-, and tensor-based fusion are often used.

Kernel-based fusion is extended based on classifiers that contain kernels, such as support vector machine (SVM). For different modalities, different kernels are used. The flexibility in kernel selection and convexity of the loss functions make multiple-kernel learning fusion popular in many applications, including MER. However, during testing, these fusion methods rely on the support vectors in the training data, which results in large memory cost and inefficient reference.

Graph-based fusion constructs separate graphs or hypergraphs for each modality, combines these graphs into a fused one, and learns the weights of different edges and modalities by graph-based learning. It can well deal with the data incompleteness problem simply by constructing graphs based on available data. Besides the extracted feature representations, we can also incorporate prior human knowledge into the models by corresponding edges. However, the computational cost would increase exponentially when more training samples are available.

Neural network-based fusion employs a direct and intuitive strategy to fuse the feature representations or predicted results of different modalities by a neural network. Attention-based fusion uses some attention mechanisms to obtain the weighted sum of a set of feature representations with scalar weights that are dynamically learned by an attention module. Different attention mechanisms correspond to fusing different components.

For example, spatial image attention measures the importance of different image regions. Image and text coattention employs symmetric attention mechanisms to generate attended visual and textual representations. Parallel and alternating coattention methods can be used to, respectively, generate attention for different modalities simultaneously and one by one.

Recently, a multimodal adaptation gate (MAG) is designed to enable transformer-based contextual word representations, such as BERT and XLNet, to accept multimodal nonverbal data [17]. Based on the attention conditioned on the nonverbal behaviors, MAG can essentially map the informative multiple modalities to a vector with a trajectory and magnitude.

Tensor-based fusion tries to exploit the correlations of different representations by some specific tensor operations, such as outer product and polynomial tensor pooling. These fusion methods for deep models are capable of learning from a large amount of data in an end-to-end manner with good performance but suffer from low interpretability.

One important property of these feature fusion methods is whether they support temporal modeling for MER in videos. It is obvious that early fusion can, while late and hybrid fusion cannot since the predicted results based on each modality are already known before late fusion. For model-based fusion, excluding kernel-based fusion, all others can be used for temporal modeling. Example methods for graph-based fusion methods include hidden Markov models (HMMs) and conditional random fields (CRFs), and RNN and LSTM networks can be employed for neural network-based fusion.

Classifier optimization for MER

For the text represented as a sequence of word embeddings, the most popular approaches to leverage the semantics among

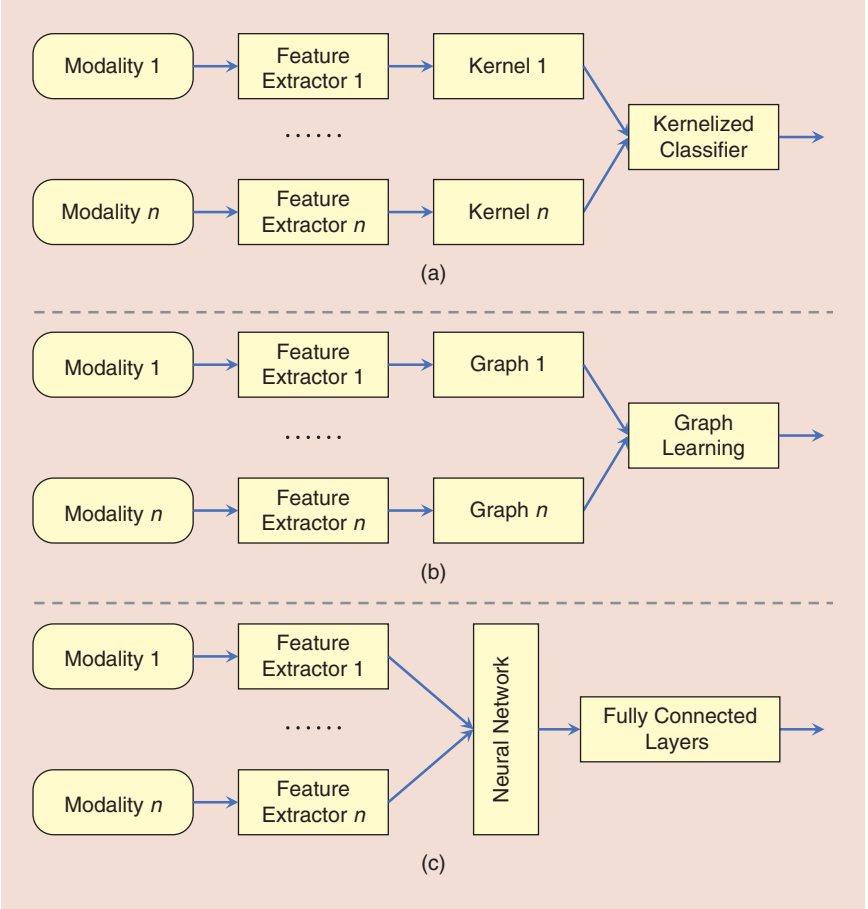


FIGURE 3. The different model-based fusion strategies, where n is the number of different modalities: (a) kernel-, (b) graph-, and (c) neural network-based fusion.

words are RNNs and CNNs. LSTM, as a typical RNN, contains a series of cells with the same structure. Every cell takes a word embedding and the hidden state from the last cell as input, computes the output, and updates the hidden state for the following cell. The hidden state records the semantics of previous words. A CNN computes local contextual features among consecutive words through convolution operations, and average- or max-pooling layers are used to further integrate the obtained features for the following sentiment classification.

Recently, researchers have begun to use transformer-based methods, e.g., BERT and GPT-3. The transformer is implemented as a series of modules containing a multihead self-attention layer followed by a normalization layer, a feed-forward network, and another normalization layer. The order of words in the text is also represented by another position embedding layer. Compared with an RNN, the transformer does not require the sequential processing of words, which improves the parallelizability, and, compared with a CNN, the transformer can model relationships between more distant words.

The classification approaches used in audio emotion recognition generally include the following two options: traditional and deep learning-based methods. For traditional methods, HMM is a representative method because of its capability of capturing the dynamic characteristics of sequential data. SVM is also widely utilized in audio emotion recognition.

Deep learning-based methods have become more popular since they are not restricted by the classical independence assumptions of HMM models. Among these techniques, sequence-to-sequence models with attention have shown success in an end-to-end manner. Recently, some approaches have significantly extended the state of the art in this area by developing deep hybrid convolutional and recurrent models [14].

In the early years, similar to this task in other modalities, multiple handcrafted image features were integrated and input into an SVM to train classifiers. Then, based on deep learning, the classifier and feature extractor were connected and optimized in an end-to-end manner by corresponding loss functions, like cross-entropy loss [26]. In addition, popular metric losses, such as triplet and *N*-pair loss, also took part in the network optimization to obtain more discriminative features.

With the described learning paradigm, each image was predicted as a single dominant emotion category. However, based on the theories of psychology, an image may evoke multiple emotions in viewers, which leads to an ambiguity problem. To address this issue, label-distribution learning is employed to predict a concrete relative degree for each emotion category, where Kullback–Leibler divergence is the most popular loss function.

Some informative and attractive regions of an image always determine the emotion of it. Therefore, a series of architectures with extra attention or detection branches is constructed. With optimization for multiple tasks, including attention and the original task, a more robust and discriminative model is obtained.

Most existing methods employ a two-stage pipeline to recognize video emotion, i.e., extracting visual and/or audio features and training classifiers. For the latter, many machine learning methods have been investigated to model the mapping between video features and discrete emotion categories, including SVM, GMM, HMM, dynamic Bayesian networks, and CRF. Although the approaches contributed to the development of emotion recognition in videos, recent methods have been proposed to recognize video emotions in an end-to-end manner based on deep neural networks due to their superior capability [27].

CNN-based methods first employ 3D CNNs to extract high-level spatiotemporal features, which contain affective information, and then use fully connected layers to classify emotions. Finally, the models are followed by the loss function to optimize the whole network. Inspired by the human process of perceiving emotions, CNN-based methods employ the attention mechanism to emphasize emotionally relevant regions of frames or segments in each video. Furthermore, considering the polar-emotion hierarchy constraint, recent methods propose polarity-consistent cross-entropy loss to guide the attention generation.

The gait of a person can be represented as a sequence of 2D or 3D joint coordinates for each frame in walking videos. To leverage the inherent affective cues in the coordinates of joints, many classifiers or architectures have been used to extract affective features in the gait. LSTM networks contain many special units, i.e., memory cells, and can store the joint coordinate information from particular time steps in a long data sequence. Thus, they were used in some early work of gait emotion recognition.

The hidden features of LSTM can be further concatenated with the handcrafted affective features and are then fed into a classifier [e.g., SVM or random forest (RF)] to predict emotions. Recently, another popular network used in gait emotion prediction is the spatial–temporal graph convolutional network (ST-GCN), which was initially proposed for action recognition from human skeletal graphs. "Spatial" represents the spatial edges in the skeletal structure, which are the limbs that connect the body joints. "Temporal" refers to temporal edges, and they connect the positions of each joint across different frames. ST-GCN can be easily implemented as a spatial followed by a temporal convolution, which is similar to deep convolutional networks.

EEG-based emotion recognition usually employs various classifiers, such as SVM, decision trees, and *k*-nearest neighbor to classify handcrafted features in the early stage. Later, since CNNs and RNNs are good at extracting the spatial and temporal information of EEG signals, respectively, end-to-end structures, such as cascade convolutional recurrent networks (which combine a CNN and RNN), LSTM-RNNs, and parallel convolutional RNNs, are successfully designed and applied to emotion recognition tasks.

Quantitative comparison of representative MER methods

To give readers an impression of the performances of state-of-the-art MER methods, we conduct experiments to fairly compare some representative methods based on the released codes

Table 2. A quantitative comparison of some representative methods for MER on five widely used data sets using GLOVE as word embeddings.

Data Set	CMU-MOSI					YouTube		ICT-MMMO		MOUD		IEMOCAP					
Train:Val:Test	1,284:229:686					30:5:11		11:2:4		49:10:20		3:1:1					
Method/Metric	A_2↑	F1↑	A_7↑	M↓	C↑	A_3↑	F1↑	A_2↑	F1↑	A_2↑	F1↑	A_9↑	F1↑	M_V↓	C_V↑	M_A↓	C_A↑
SVM	71.6	72.3	26.5	1.1	0.559	42.4	37.9	68.8	68.7	60.4	45.5	24.1	18	0.251	0.06	0.546	0.54
RF	56.4	56.3	21.3	—	—	49.3	49.2	70	69.8	64.2	63.3	27.3	25.3	—	—	—	—
THMM	50.7	45.4	17.8	—	—	42.4	27.9	53.8	53	58.5	52.7	23.5	10.8	—	—	—	—
MV-LSTM	73.9	74	33.2	1.019	0.601	45.8	43.3	72.5	72.3	57.6	48.2	31.3	26.7	0.257	0.02	0.513	0.62
BC-LSTM	73.9	73.9	28.7	1.079	0.581	47.5	47.3	70	71.1	72.6	72.9	35.9	34.1	0.248	0.07	0.593	0.4
TFN	74.6	74.5	28.7	1.04	0.587	47.5	41	72.5	72.6	63.2	61.7	36	34.5	0.251	0.04	0.521	0.55
MARN	77.1	77	34.7	0.968	0.625	54.2	52.9	86.3	85.9	81.1	81.2	37	35.9	0.242	0.1	0.497	0.65
MFN	77.4	77.3	34.1	0.965	0.632	61	60.7	87.5	87.1	81.1	80.4	36.5	34.9	0.236	0.111	0.482	0.645

A_N and F1 are percentages. ↑ and ↓ respectively indicate that higher and lower values represent better performance for corresponding metrics (the same below).
Evaluation metrics: A_N: emotion classification accuracy, where N denotes the number of emotion classes; M: mean absolute error; C: Pearson correlation; V: valence results; A: arousal results.

of the CMU multimodal software development kit [37] and MAG [38]. Specifically, the compared nondeep methods include SVM, RF, and trimodal HMM (THMM); the compared deep methods include multiview LSTM (MV-LSTM), bidirectional contextual LSTM (BC-LSTM), tensor fusion network (TFN), multiattention recurrent network (MARN), memory fusion network (MFN), fine-tuning (FT), and MAG.

We conduct experiments on five data sets: CMU-MOSI, YouTube, ICT-MMMO, MOUD, and IEMOCAP. All of the data sets contain three modalities: face, speech, and transcript text. For visual features, Facet is used to extract per-frame basic and advanced emotions and facial action units as indicators of facial muscle movement. For acoustic features, COVAREP is employed to extract 12 mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters, and maxima dispersion quotients. For linguistic features, three different pretrained word embeddings, i.e., GLOVE, BERT, and XLNet, are employed to obtain the word vector. For comparison, the human performance is also reported on CMU-MOSI with results derived from [39].

The input to the nondeep methods is the early fusion of multimodal features. For emotion classification, we use accuracy and F1 as metrics; for emotion regression, we use mean absolute error and the Pearson correlation. Higher values indicate better performance for all of the metrics, except mean absolute error, where lower values denote better performance.

From the results in Tables 2 and 3, we have the following observations. First, the performances of deep models are generally better than nondeep ones. Second, for different data sets, the methods with the best performances are different. For example, RF achieves the best performance among nondeep models except CMU-MOSI, which demonstrates its good generalization ability, while the performance of SVM is much better than that of RF or THMM on CMU-MOSI.

Third, multiclass classification is more difficult than binary classification, such as 77.1 versus 34.7 of MARN on CMU-MOSI. Fourth, comparing the same method in Tables 2 and 3 on CMU-MOSI, we can conclude that BERT and XLNet can provide better word embeddings than GLOVE, and XLNet is generally better than BERT. Finally, although XLNet-based MAG achieves a near-human-level performance on CMU-MOSI, there is still some gap, and more efforts are expected to achieve even better performance than humans.

Domain adaptation for MER

Domain adaptation aims to learn a transferable MER model from labeled source domains that can perform well on unlabeled target domains [23]. Recent efforts have been dedicated to deep unsupervised domain adaptation [23], which employs a two-stream architecture. One stream is used to train an MER model on the labeled source domains, while the other is used to align the source and target domains. Based on the alignment strategy, existing unimodal domain adaptation approaches can be classified into different categories [23], such as discrepancy-based, adversarial discriminative, adversarial generative, and self-supervision-based methods.

Discrepancy-based methods employ some distance metrics to explicitly measure the discrepancy between the source and target domains on the corresponding activation layers of the two network streams. Commonly used discrepancy loss measures include maximum mean discrepancy, correlation alignment, geodesic distance, central moment discrepancy, Wasserstein discrepancy, contrastive Domain discrepancy, and higher-order

Table 3. A quantitative comparison of some representative methods for MER on the CMU-MOSI data set using BERT or XLNet as word embeddings.

Method/Metric	A_2↑	F1↑	M↓	C↑
TFN	74.8/78.2	74.1/78.2	0.955/0.914	0.649/0.713
MARN	77.7/78.3	77.9/78.8	0.938/0.921	0.691/0.707
MFN	78.2/78.3	78.1/78.4	0.911/0.898	0.699/0.713
FT	83.5/84.7	83.4/84.6	0.739/0.676	0.782/0.812
MAG	84.2/85.7	84.1/85.6	0.712/0.675	0.796/0.821
Human	85.7	87.5	0.71	0.82

The numbers on the left and right sides of "/" are the MER results based on BERT and XLNet, respectively.

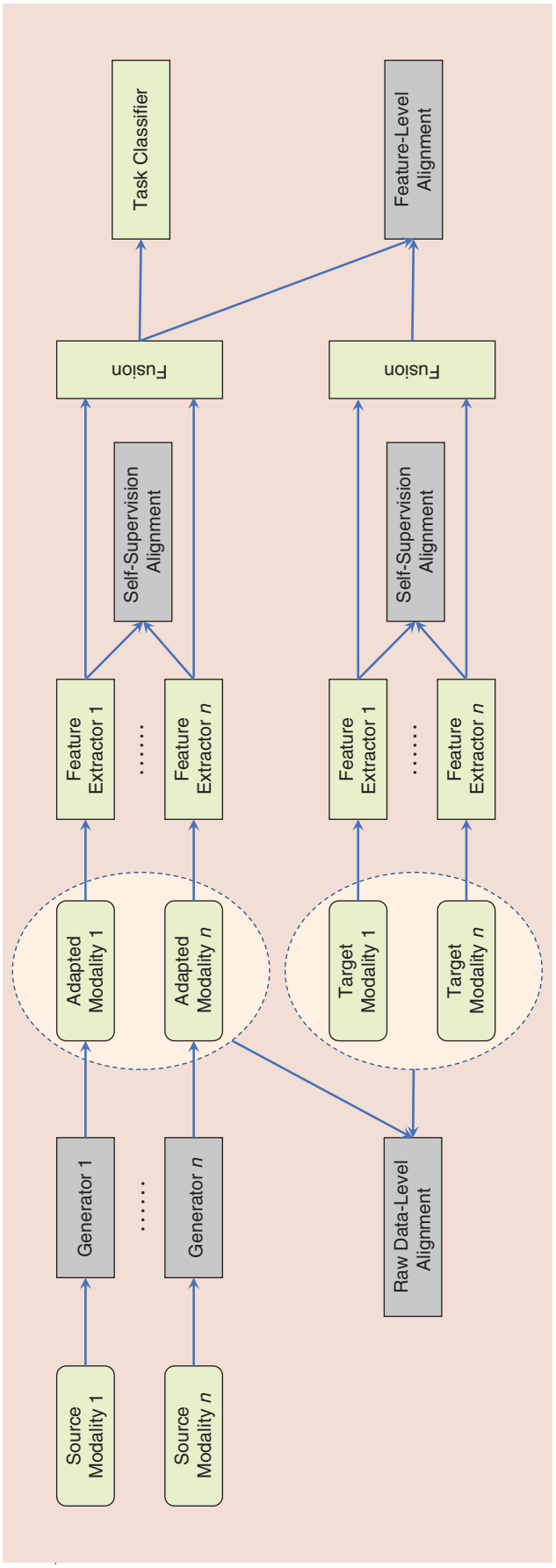


FIGURE 4. A generalized framework for multimodal domain adaptation with one labeled source domain and one unlabeled target domain. The grayscale rectangles with text in bold represent different alignment strategies. Most existing multimodal domain adaptation methods can be obtained by employing different component details, enforcing some constraints, or slightly changing the architecture.

moment matching. Besides the used discrepancy loss, there are some other differences among existing methods, such as whether the loss is at the domain or class level, which layer the loss is operated on, whether the backbone networks share weights or not, and if the aligned distribution is marginal or joint.

Adversarial discriminative models usually align the source and target domains with a domain discriminator by adversarially making different domains indistinguishable. The input to the discriminator ranges from the original data to extracted features, and the adversarial alignment can be global or classwise. We can also consider using shared or unshared feature extractors.

Adversarial generative models usually employ a generator to create fake source or target data to make the domain discriminator indistinguishable from the generated and real domains. The generator is typically based on a generative adversarial network (GAN) and its variants, such as CoGAN, SimGAN, and CycleGAN. The input to the generator and discriminator can be different in different methods.

Self-supervision-based methods combine some auxiliary self-supervised learning tasks, such as reconstruction, image rotation prediction, jigsaw prediction, and masking, with the original task network to bring the source and target domains closer. We can compare these four types of domain adaptation methods from the perspectives of theory guarantee, efficiency, task scalability, data scalability, data dependency, optimizability, and performance. We can combine some of these techniques to jointly exploit their advantages.

The main difficulty in domain adaptation for MER lies in the alignment of multiple modalities between the source and target domains simultaneously. There are some simple but effective ways to extend unimodal domain adaptation to multimodal settings, as shown in Figure 4. For example, we can use discrepancy loss or a discriminator to align the fused feature representations. The correspondence between different modalities can be used as a self-supervised alignment.

Extending adversarial generative models from unimodal to multimodal would be more difficult. Unlike images, other generated modalities, such as text and speech, might have confused semantics, although they can make the discriminator indistinguishable. Generating intermediate feature representations instead of raw data can provide a feasible solution.

Applications

Recognizing emotions from multiple explicit cues and implicit stimuli is of great significance in a broad range of real-world applications. Generally speaking, emotion is the most important aspect of the quality and meaning of our existence, and it makes life worth living. The emotional impact of digital data lies in that it can improve the user experience of existing techniques and then strengthen the knowledge transfer between people and computers [18].

Many people tend to post texts, images, and videos on social networks to express their daily feelings about life. Inspired by this, we can mine people's opinions and sentiments toward topics and events happening in the real world [28]. For instance,

user-generated content on Facebook or Instagram can be used to derive the attitudes of people from different countries and regions when they face a pandemic like COVID-19 [29]. Researchers also try to detect sentiment in social networks and apply the results to predict political elections. Note that, when the personalized emotion of an individual is detected, we can further group these emotions, which may contribute to predicting the tendencies of society.

Another important application of MER is business intelligence, especially marketing and consumer behavior analysis [30]. Today, most apparel e-retailers use human models to present products. The model's face presentation is proven to have a significant effect on consumer approach behavior. To be specific, for participants whose emotional receptivity is high, a smiling facial expression tends to lead to the highest approach behavior. In addition, researchers examine how online store specialization influences consumer pleasure and arousal, based on the stimulus–organism–response framework.

Emotion recognition can also be used in call centers, the goal of which is to detect the emotional states of both the caller and operator. The system recognizes the involved emotions through the intonation and tempo as well as the texts translated from the corresponding speech. Based on this, we can receive feedback on the quality of the service.

Meanwhile, emotion recognition plays an important role in the field of medical treatment and psychological health. With the popularity of social media, some people prefer expressing their emotions over the Internet rather than to others. If a user is observed to be sharing negative information (e.g., sadness) frequently and continuously, it is necessary to track her or his mental status to prevent the occurrence of psychological illness and even suicide.

Emotional states can also be used to monitor and predict the fatigue level of a variety of people, like drivers, pilots, workers on assembly lines, and students in classrooms. This technique both prevents dangerous situations and benefits the evaluation of work/study efficiency. Further, emotional states can be incorporated into various security applications, such as systems for monitoring public spaces (e.g., bus/train/subway stations or football stadiums) for potential aggression.

Recently, an effective auxiliary system was introduced in the diagnosis and treatment process of autism spectrum disorder (ASD) in children to assist in collecting information on the condition. To help professional clinicians better and faster make a diagnosis and give treatment to ASD patients, this system characterizes facial expressions and eye gaze attention, which are considered to be remarkable indicators for the early screening of autism.

MER is used to improve the personal entertainment experience. For example, a recent work in the brainwave–music interface maps EEG characteristics to musical structures (note, intensity, and pitch). Similarly, efforts have been made to understand the emotion-centric correlation between different modalities that are essential for various applications. Affective image–music matching provides a good chance appending a sequence of music to a given image such that they both evoke the same emotion. This helps generate emotion-aware music playlists from one's personal album photos in mobile devices.

Future directions

Existing methods have achieved promising performances in various MER settings, such as visual–audio, facial–textual–speech, and textual–visual tasks. However, all of the summarized challenges have not been fully addressed. For example, the issues of how to extract discriminative features that are more related to emotion, balance common and personalized emotion reactions, and emphasize the more important modalities are still open. To help improve the performances of MER methods and make them fit special requirements in the real world, we provide some potential future directions.

New methodologies for MER

- *Contextual and prior knowledge modeling*: The experienced emotion of a user can be significantly influenced by contextual information, such as the conversational and social environments. The prior knowledge of users, such as personality and age, can also contribute to emotion perception. For example, an optimistic user and a pessimistic viewer are likely to see different aspects of the same stimuli. Jointly considering this important contextual information and prior knowledge is expected to improve the MER performance. Graph-related methods, such as graph convolutional networks, are possible solutions to model the relationships among factors and emotions.
- *Learning from unlabeled, unreliable, and unmatched affective signals*: In the big data era, the affective data might be sparsely labeled or even unlabeled, raw data or labels can be unreliable, and test and training data could be unmatched. Exploring advanced machine learning techniques, such as unsupervised representation learning, dynamic data selection and balancing, and domain adaptation, as well as the embedding of special properties of emotions, can help to address these challenges.
- *Explainable, robust, and secure deep learning for MER*: Due to the black-box nature, it is difficult to understand why existing deep neural networks perform well for MER, and the trained deep networks are vulnerable to adversarial attacks and inevitable noises that might cause erraticism. Essentially explaining the decision-making process of deep learning can help with the design of robust and secure MER systems.
- *A combination of explicit and implicit signals*: Both explicit and implicit signals are demonstrated to be useful for MER, but they also suffer from some limitations. For example, explicit signals can be easily suppressed or are difficult to capture, while implicit signals might not reflect the emotions in real time. Jointly combining them to explore complementary information during a viewer–multimedia interaction would boost the MER performance.
- *The incorporation of emotion theory into MER*: Different theories have been proposed in psychology, physiology, neurology, and the cognitive sciences. These theories can help us understand how humans produce emotion, but they

have not been employed in the computational MER task. We believe such an incorporation would make more sense to recognize emotions.

More practical MER settings

- *MER in the wild*: Current MER methods mainly focus on neat lab settings. However, MER problems in the real world are much more complex. For example, the collected data might contain much noise that is unrelated to emotion; the users in the test set could be from different cultures and languages from those in the training set, resulting in varying ways of expressing emotion; different emotion label spaces might be employed across various settings; or training data may be incrementally available. Designing an effective MER model that is generalizable to these practical settings is worth investigating.
- *MER on the edge*: When deploying MER models in edge devices, such as mobile phones and security cameras, we have to consider the computing limitations and data privacy. Techniques like autopruning, neural architecture search, invertible neural network, and software–hardware co-design are believed to be beneficial for efficient on-device training.
- *Personalized and group MER*: Because of the subjectivity of emotions, simply recognizing the dominant emotion of different individuals is insufficient. It is ideal but impractical to collect enough data for each individual to train personalized MER models. Adapting the well-trained MER models for dominant emotions to each individual with a small amount of labeled data is a possible alternate solution. On the other hand, it would make more sense to predict emotions for groups of individuals who share similar tastes or interests and have a similar background. Group emotion recognition is essential in many applications, such as recommendation systems, but how to classify users into different groups is still challenging.

Real applications based on MER

- *The implementation of MER in real-world applications*: Although emotion recognition has been emphasized to be important for decades, it has rarely been applied to real scenarios due to relatively low performance. With the recent rapid progress of MER, we can begin incorporating emotion into different applications in the marketing, education, health care, and service sectors. The feedback from the applications can, in turn, promote the development of MER. Together with emotion generation, we believe an age of artificial emotional intelligence is coming.
- *Wearable, simple, and accurate affective data collection*: To conduct MER tasks, the first step is to collect accurate affective data. Developing wearable, simple, and even contactless sensors to capture such data would make it more acceptable to users.
- *Security, privacy, ethics, and fairness of MER*: During data collection, it is possible to extract users' confidential information, such as identity, age, and so on. Protecting the security and privacy of users and avoiding any chance of misuse must be taken into consideration. Emotion recognition in real applications might have a negative and even dangerous impact on a person, such as emotional pressure. Methods to eliminate such an effect should also be considered from the perspectives of ethics and fairness.

Conclusions

In this article, we provided a comprehensive tutorial on MER. We briefly introduced emotion representation models, both explicit and implicit affective modalities, emotion annotations, and corresponding computational tasks. We summarized the main challenges of MER in detail, and then we emphatically introduced different computational methodologies, including the representation learning of each affective modality, feature fusion of different affective modalities, and classifier optimization as well as domain adaptation for MER. We ended this tutorial with discussions of real-world applications and future directions. We hope this tutorial can motivate novel techniques to facilitate the development of MER, and we believe that this area will continue to attract significant research efforts.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (grant 2018AAA0100403), National Natural Science Foundation of China (grants 61701273, 61876094, U1933114, 61925107, and U1936202), Natural Science Foundation of Tianjin, China (grants 20JCJQJC00020, 18JCYBJC15400, and 18ZXZNGX00110), and Berkeley DeepDrive. Jufeng Yang is the corresponding author of this article.

Authors

Sicheng Zhao (schzhao@gmail.com) received his Ph.D. degree from the Harbin Institute of Technology, Harbin, China, in 2016. He is a postdoctoral research scientist at Columbia University, New York, New York, 10032, USA. He was a visiting scholar at the National University of Singapore from July 2013 to June 2014, a research fellow at Tsinghua University from September 2016 to September 2017, and a research fellow at the University of California, Berkeley from September 2017 to September 2020. His research interests include affective computing, multimedia, and computer vision. He is a Senior Member of IEEE.

Guoli Jia (exped1230@gmail.com) is working toward his master's degree at the College of Computer Science, Nankai University, Tianjin, 300350, China. His research interests include computer vision and pattern recognition.

Jufeng Yang (yangjufeng@nankai.edu.cn) received his Ph.D. degree from Nankai University, Tianjin, China, in 2009. He is a full professor in the College of Computer Science, Nankai University, Tianjin, 300350, China, and was a visiting scholar with the Vision and Learning Lab, University of California, Merced, USA, from 2015 to 2016. His recent interests include computer vision, machine learning, and multimedia. He is a Member of IEEE.

Guiguang Ding (dinggg@tsinghua.edu.cn) received his Ph.D. degree from Xidian University, China, in 2004. He is a full professor with School of Software, Tsinghua University, Beijing, 100084, China. Before joining the School of Software

in 2006, he was a postdoctoral research fellow in the Department of Automation, Tsinghua University. He was a leading guest editor of *Neural Processing Letters* and *Multimedia Tools and Applications*. He served as special session chair of the 2021 IEEE ICASSP; 2019 and 2020 IEEE ICME; and 2017 Pacific Rim Conference on Multimedia and reviewer for more than 20 prestigious international journals and conferences. His research interests include the area of multimedia information retrieval, computer vision, and machine learning. He is a Member of IEEE.

Kurt Keutzer (keutzer@berkeley.edu) received his Ph.D. degree in computer science from Indiana University in 1984. He then joined the research division of AT&T Bell Laboratories. In 1991, he joined Synopsys, Inc., where he ultimately became CTO and senior vice-president of research. In 1998, he became a professor of electrical engineering and computer science at the University of California, Berkeley, Berkeley, California, 94720, USA. He has published six books and more than 250 refereed articles, and he is among the most highly cited authors in hardware and design automation. His research interests include using parallelism to accelerate the training and deployment of deep neural networks for applications in computer vision, speech recognition, multimedia analysis, and computational finance. He is a Life Fellow of IEEE.

References

[1] D. Kahneman, *Thinking, Fast and Slow*. New York: Macmillan, 2011.

[2] M. Minsky, *The Society of Mind*. New York: Simon and Schuster, 1986.

[3] D. Schuller and B. W. Schuller, "The age of artificial emotional intelligence," *Computer*, vol. 51, no. 9, pp. 38–46, 2018. doi: 10.1109/MC.2018.3620963.

[4] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, Sept. 2017. doi: 10.1016/j.imavis.2017.08.003.

[5] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim, "Exploring fusion methods for multimodal emotion recognition with missing data," *IEEE Trans. Affective Comput.*, vol. 2, no. 4, pp. 206–218, 2011. doi: 10.1109/T-AFFC.2011.12.

[6] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, p. 43, 2015. doi: 10.1145/2682899.

[7] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, 2017. doi: 10.1109/MSP.2017.2738401.

[8] S. K. D'Mello, N. Bosch, and H. Chen, "Multimodal-multisensor affect detection," in *Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, Detection Emotion Cognition*, vol. 2. New York: Association for Computing Machinery and Morgan & Claypool 2018, pp. 167–202.

[9] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

[10] S. Zhao, X. Yao, J. Yang, G. Jia, G. Ding, T.-S. Chua, B. W. Schuller, and K. Keutzer, "Affective image content analysis: Two decades review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. doi: 10.1109/TPAMI.2021.3094362.

[11] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affective Comput.*, vol. 5, no. 2, pp. 101–111, 2014. doi: 10.1109/TAFFC.2014.2317187.

[12] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009. doi: 10.1109/TPAMI.2008.52.

[13] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 107–129, 2017. doi: 10.1109/MSP.2017.2699358.

[14] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020. doi: 10.1016/j.specom.2019.12.001.

[15] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "Ascertain: Emotion and personality recognition using commercial sensors," *IEEE Trans. Affective Comput.*, vol. 9, no. 2, pp. 147–160, 2018. doi: 10.1109/TAFFC.2016.2625250.

[16] A. Giachanou and F. Crestani, "Like it or not: A survey of twitter sentiment analysis methods," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1–41, 2016. doi: 10.1145/2938640.

[17] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.

[18] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Process. Mag.*, vol. 28, no. 5, pp. 94–115, 2011. doi: 10.1109/MSP.2011.941851.

[19] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affective Comput.*, vol. 6, no. 4, pp. 410–430, 2015. doi: 10.1109/TAFFC.2015.2432791.

[20] J. Yang, M. Sun, and S. Xiaoxiao, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 224–230.

[21] S. Zhao, H. Yao, Y. Gao, R. Ji, and G. Ding, "Continuous probability distribution prediction of image emotions via multitask shared sparse regression," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 632–645, 2017. doi: 10.1109/TMM.2016.2617741.

[22] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2020. doi: 10.1109/TMM.2019.2939744.

[23] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J. E. Gonzalez et al., "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, 2020. doi: 10.1109/TNNLS.2020.3028503.

[24] U. Bhattacharya, T. Mittal, R. Chandra, T. Randhavane, A. Bera, and D. Manocha, "Step: Spatial temporal graph convolutional networks for emotion perception from gaits," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 1342–1350. doi: 10.1609/aaai.v34i02.5490.

[25] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affective Comput.*, vol. 9, no. 4, pp. 550–562, 2018. doi: 10.1109/TAFFC.2017.2660485.

[26] A. Hu and S. Flaxman, "Multimodal sentiment analysis to explore the structure of emotions," in *Proc. ACM Int. Conf. Knowledge Discovery Data Mining*, 2018, pp. 350–358. doi: 10.1145/3219819.3219853.

[27] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interaction*, 2015, pp. 467–474. doi: 10.1145/2818346.2830596.

[28] R. Ji, F. Chen, L. Cao, and Y. Gao, "Cross-modality microblog sentiment prediction via bi-layer multimodal hypergraph learning," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1062–1075, 2019. doi: 10.1109/TMM.2018.2867718.

[29] H. Lyu, L. Chen, Y. Wang, and J. Luo, "Sense and sensibility: Characterizing social media users regarding the use of controversial terms for COVID-19," *IEEE Trans. Big Data*, to be published. doi: 10.1109/TBDATA.2020.2996401.

[30] R. Wu and C. L. Wang, "The asymmetric impact of other-blame regret versus self-blame regret on negative word of mouth: Empirical evidence from China," *European J. Marketing*, vol. 51, no. 11/12, pp. 1799–1816, 2017. doi: 10.1108/EJM-06-2015-0322.

[31] K. Diehl, A. C. Morales, G. J. Fitzsimmons, and D. Simester, "Shopping interdependencies: How emotions affect consumer search and shopping behavior." [Online]. Available: https://msbfile03.usc.edu/digitalmeasures/kdiehl/intellcont/Shopping%20Interdependencies%20WP-1.pdf

[32] T. A. Dingus, F. Guo, S. Lee, J. F. Antin, M. Perez, M. Buchanan-King, and J. Hankey, "Driver crash risk factors and prevalence evaluation using naturalistic driving data," *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 10, pp. 2636–2641, 2016. [Online]. Available: https://www.pnas.org/content/113/10/2636

[33] K. Trezise, A. Bourgeois, and C. Luck, "Emotions in classrooms: The need to understand how emotions affect learning and education," npj Science of Learning Community, July 13, 2017. [Online]. Available: https://npjscilearncommunity.nature.com/posts/18507

[34] F. M. Marchak, "Detecting false intent using eye blink measures," *Front. Psychol.*, vol. 4, p. 736, Oct. 2013. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00736/full

[35] H. K. M. Meeren, C. C. R. J. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 45, pp. 16,518–16,523, 2016. [Online]. Available: https://www.pnas.org/content/102/45/16518.short

[36] P. Chakravorty, "What Is a Signal? [Lecture Notes]," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 175–177, Sept. 2018. doi: 10.1109/MSP.2018.2832195

[37] "A2Zadeh/CMU-MultimodalSDK," GitHub. https://github.com/A2Zadeh/CMU-MultimodalSDK (accessed Sept. 3, 2021).

[38] https://github.com/WasifurRahman/BERT_multimodal_transformer

[39] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proc. 32nd AAAI Conf. Artf. Intell.*, 2018, pp. 5642–5649.

SP