

Unlocking the Emotional World of Visual Media: An Overview of the Science, Research, and Impact of Understanding Emotion

Drawing insights from psychology, engineering, and the arts, this article provides a comprehensive overview of the field of emotion analysis in visual media, and discusses the latest research, systems, challenges, ethical implications, and potential impact of artificial emotional intelligence on society.

By James Z. Wang, *Senior Member, IEEE*, Sicheng Zhao, *Senior Member, IEEE*, Chenyan Wu, Reginald B. Adams, Jr., Michelle G. Newman, Tal Shafir, Rachelle Tsachor

May 24, 2023

Abstract—The emergence of artificial emotional intelligence technology is revolutionizing the fields of computers and robotics, allowing for a new level of communication and understanding of human behavior that was once thought impossible. Whereas recent advancements in deep learning have transformed the field of computer vision, automated understanding of evoked or expressed emotions in visual media remains in its infancy. This founding stems from the absence of a universally accepted definition of “emotion,” coupled with the inherently subjective nature of emotions and their intricate nuances. In this paper, we provide a comprehensive, multidisciplinary overview of the field of emotion analysis in visual media, drawing on insights from psychology, engineering, and the arts. We begin by exploring the psychological foundations of emotion and the computational principles that underpin the understanding of emotions from images and videos. We then review the latest research and systems within the field, accentuating the most promising approaches. We also discuss the current technological challenges and limitations of emotion analysis, underscoring the necessity for continued investigation and innovation. We contend that this represents a “Holy Grail” research problem in computing and delineate pivotal directions for future inquiry. Finally, we examine the ethical ramifications of emotion-understanding technologies and contemplate their potential societal impacts. Overall, this article endeavors to equip readers with a deeper understanding of the domain of emotion analysis in visual media and to inspire further

research and development in this captivating and rapidly evolving field.

Index Terms—Evoked emotion; expressed emotion; bodily expressed emotion understanding; psychology; movement analysis; artificial emotional intelligence; intelligent robots; deep learning; human behavior; ethics.

I. INTRODUCTION

As Artificial Intelligence (AI) technology becomes more prevalent and capable of performing a wide range of tasks, the need for effective communication between humans and AI systems is becoming increasingly important. The adoption of smart home products and services is projected to reach 400 million worldwide, with smart devices such as Alexa and Astro becoming increasingly common in households [1]. However, these devices are currently limited to executing specific commands and do not possess the capability to understand or respond to human emotions [2]. This lack of emotional intelligence (EQ) limits their potential applications, and this constraint is particularly relevant for future robotic applications, such as personal assistant robots, social robots, service robots, factory/warehouse robots, and police robots, which require close collaboration and a comprehensive understanding of human behavior and emotions.

The ability to impart EQ to AI when dealing with visual information is a topic of growing interest. This article aims to address the fundamental question of *how to “teach” AI to understand and respond to human emotions based on images and videos*. The potential technical solutions to these questions have far-reaching implications for various application domains, including human-AI interaction, autonomous driving, social media, entertainment, information management and retrieval, design, industrial safety, and education.

To provide a comprehensive and well-balanced view of this complex subject, it is essential to draw on the expertise of various fields, including computer and information science and engineering, psychology, data science, movement analysis, and performing arts. The interdisciplinary nature of this topic highlights the need for collaboration and cooperation among researchers from different fields in order to achieve a deeper understanding of the subject.

*Manuscript received 31 August 2022; revised 7 March 2023 and 15 April 2023; accepted 3 May 2023. The work of James Z. Wang and Chenyan Wu was supported in part by generous gifts from the Amazon Research Awards Program. The work of James Z. Wang, Reginald B. Adams Jr., and Michelle G. Newman was supported in part by the National Science Foundation (NSF) under Grant IIS-1110970 and Grant CNS-1921783. The work of James Z. Wang, Reginald B. Adams Jr., Michelle G. Newman, Tal Shafir, and Rachelle Tsachor was supported in part by the NSF under Grant CNS-2234195 and Grant CNS-2234197. The work of James Z. Wang and Michelle G. Newman was supported in part by the NSF under Grant CIF-2205004. (*Corresponding author: James Z. Wang.*)

James Z. Wang and **Chenyan Wu** are with the Data Science and Artificial Intelligence Area and the Human-Computer Interaction Area of the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA (e-mail: jwang@psu.edu; czw390@psu.edu)

Reginald B. Adams, Jr. and **Michelle G. Newman** are with the Department of Psychology, The Pennsylvania State University, University Park, PA 16802, USA (e-mail: rba10@psu.edu; mgn1@psu.edu)

Sicheng Zhao is with BNRist, Tsinghua University, Beijing 100084, China. (e-mail: schzhao@tsinghua.edu.cn)

Tal Shafir is with the Emily Sagol Creative Arts Therapies Research Center, University of Haifa, 3498838, Israel (e-mail: tshafir1@univ.haifa.ac.il)

Rachelle Tsachor is with the School of Theatre and Music, University of Illinois, Chicago, IL 60607, USA (e-mail: rtsachor@uic.edu)

In this article, we focus on the topic of affective visual information analysis as it represents a highly nuanced and complex area of study with strong connections to well-established scholarly fields such as computer vision, multimedia, and image and video processing. However, it is important to note that the techniques presented here can be integrated with other data modalities, such as speech, sensor-generated streaming data, and text, in order to enhance the performance of real-world applications.

The primary objective of this article is to introduce the technical communities to the emerging field of affective visual information analysis. Recognizing the breadth and dynamic nature of this field, we do not aim to provide a comprehensive survey of all sub-areas. Instead, our discussion focuses on the fundamental psychological and computational principles (Sections II and III), recent advancements and developments (Section IV), core challenges and open issues (Section V), connections to other areas of research and development (Section VI), and ethical considerations related to this new technology (Section VII). We apologize in advance for any important publications that may have been omitted in our discussion.

Recently, there have been some other surveys and reviews on artificial emotional intelligence, such as facial expression recognition [3, 4, 5, 6], microexpression recognition [7, 8, 9], textual sentiment classification [10, 11, 12], music and speech emotion recognition [13, 14, 15], affective image content analysis [16], emotional body gesture recognition [17], bodily expressed emotion recognition [18], emotion recognition from physiological signals [19, 20], multimodal emotion recognition [21, 22], and affective theory use [23]. These articles mainly focus on emotion and sentiment analysis for a specific modality from the perspective of machine learning and pattern recognition or focus on the psychological emotion theories. Cambria [24] summarized the common tasks of affective computing and sentiment analysis and classified existing methods into three main categories: knowledge-based, statistical, and hybrid approaches. Poria et al. [25] and Wang et al. [26] reviewed both unimodal and multimodal emotion recognition before the year of 2017 and between the years of 2017 and 2020, respectively. As opposed to those reviews, the current paper aims to provide a comprehensive overview of emotion analysis from visual media (e.g., both images and videos) with insights drawn from multiple disciplines.

II. EMOTION: THE PSYCHOLOGICAL FOUNDATION

How we define emotion largely descends from the theoretical framework used to study it. In this section, we provide an overview of the most prominent emotion theories, beginning with Darwin, and underscore how contemporary dimensional approaches to understanding emotion align with both the processing of emotion by the human brain and current computer vision approaches for modeling emotion to make predictions about human perception (Section II-A). In addition, we examine the intrinsic link between emotion and adaptive behavior, a contention that is largely shared across different emotion theories (Section II-B).

A. Definitions and Models of Emotion

One of the first emotion theories put forth was Charles Darwin's in his seminal book, "On the Expression of the Emotions in Man and Animals" [32]. This book proposed that humans possess a finite set of biologically privileged emotions that evolved to confer upon us survival-related behavior. William James [33] later added to this arguing that the experience of emotion is ultimately our experience of distinct patterns of physiological arousal and physical behaviors associated with each emotion [34]. Building upon these assumptions, Ekman's Neurocultural Theory of Emotion [35] further perpetuated the notion that there exists a "universal affect program" that underlies the experience and expression of several discrete emotions, such as anger, fear, sadness, and happiness. According to this theory, basic emotional experiences and emotional displays evolved as adaptive responses to specific environmental contingencies, and thus felt, expressed, and recognized emotions are uniform across all people and cultures, and are marked by specific patterns of physiological and neural responsivity.

Considerable research has since questioned the utility of this approach. This includes findings that people: 1) are often ill-equipped to describe their own emotions in discrete emotion terms, both in research and clinical settings [36], 2) show low consensus in their ability to categorize both facial and vocal expressions of emotions in discrete emotion terms [37], and 3) show high intercorrelations across the emotional experiences they do report [38]. Such findings have prompted many researchers to explore alternative approaches to conceptualizing and measuring emotional experience that necessarily involve a cognitive component.

Magda Arnold's cognitive appraisal theory of emotion was the first to introduce the necessity of cognition in emotion elicitation [39]. Although she did not disagree with Darwin and James that emotions are adaptive states spurring on survival-related behavior, she nonetheless took them to task for not considering vast individual variation in emotional experiences. Arnold rightly underscored the capacity for the same emotion-evoking events to lead to different emotional experiences in different people. This becomes readily apparent when considering one's own emotional experiences. For example, diving off a cliff may generate an aversive fear state in one person but an enjoyable thrill state in another. The difference in how an event is evaluated, therefore, shapes the emotion that results. Central to her theory was the importance of cognitive appraisal in initially eliciting an emotion. Once elicited, she largely agreed with Darwin's functional assumptions regarding the survival benefits of emotion-related behavior. Later, Schachter and Singer [40] drew on these insights to help resolve ongoing debates regarding James' theory of embodied emotional experience. Their research demonstrated that the emotion we experience when adrenaline is released in our body depends on our cognitive framing and context. Those injected with adrenaline reported feeling happier when in a fun context and more irritated when in an angering context. The only difference was the cognitive appraisal that framed the experience of that arousal.

Building upon these ideas further, Mehrabian and Rus-

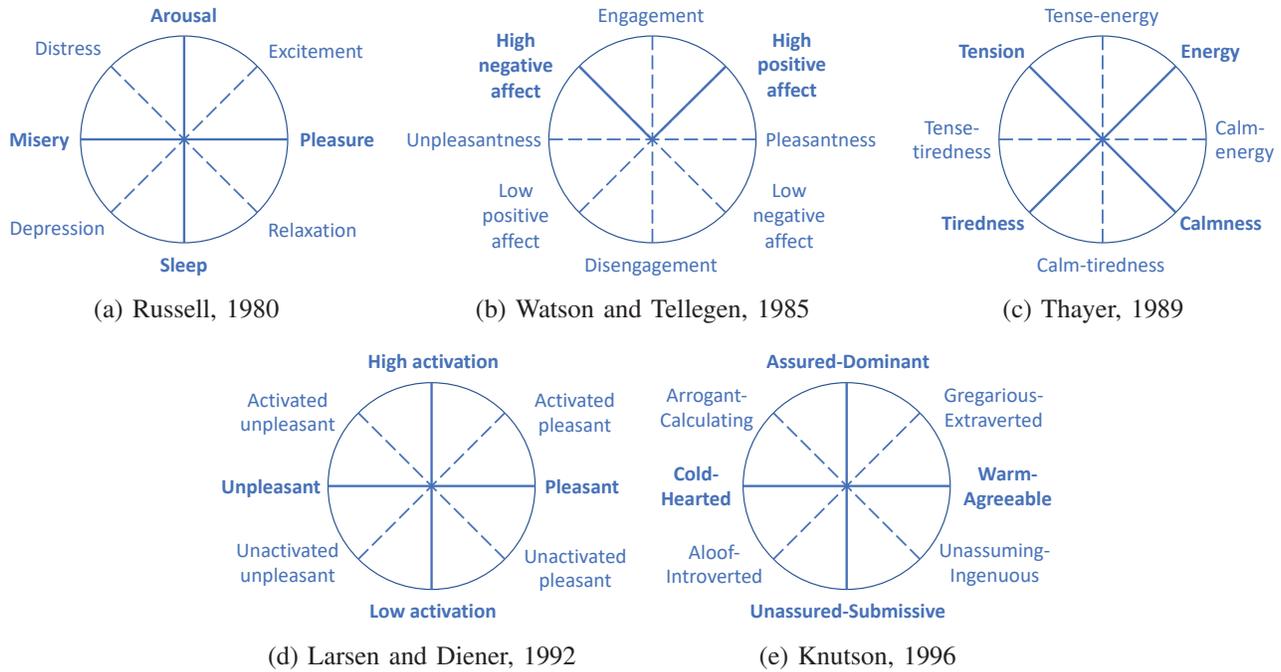


Fig. 1. Five influential circumplex emotion models developed in the psychology literature [27, 28, 29, 30, 31].

sell [41] proposed the Pleasure, Arousal, and Dominance (PAD) Emotional State Model, which suggests a dimensional account of emotion, one in which PAD constitutes the fundamental dimensions that represent all emotions. Later, Russell dropped dominance as a key dimension and focused on what he refers to as “core affect,” suggesting that all emotions can be reduced to the fundamental psychological and biological dimensions of pleasantness and arousal [42]. Dimensional approaches offer a way of conceptualizing and assessing emotion that closely approximates how the human brain processes emotion (see Fig. 1 for a comparison of different dimensional/circumplex models).

Notably, some early attempts to use computational methods to predict human emotions elicited by visual scenes employed the discrete emotion approach described at the outset of this section. For example, Mikels et al. [43] and Machajdik and Hanbury [44] used categorical approaches to assess the visual properties of stimuli taken from the International Affective Picture System (IAPS) [45], a widely used set of emotionally evocative photographs in the emotion literature. However, such approaches resulted in high levels of multicollinearity between emotions, making it difficult to disentangle emotions using traditional regression models. In contrast, adopting a dimensional approach not only aligns well with emerging theoretical accounts of emotion, but has been validated by James Wang and his colleagues in the successful assessment of human aesthetics and emotions evoked by visual scenes, as well as bodily expressed emotion [46, 47, 48, 49, 50, 51, 52]. This offers a methodological approach that is consistent with dimensional theories of emotion.

B. The Interplay Between Emotion and Behavior

Fridlund’s behavioral ecology perspective of emotion argues that emotional expression evolved primarily as a means of signaling behavioral intent [53]. The view that facial expression evolved specifically as a way to forecast behavioral intentions and consequences to others drew from Darwin’s seminal writings on expression [32], even though Darwin himself argued that expressions did not evolve for social communication per se. Fridlund’s argument is based on the idea that perceiving behavioral intentions is adaptive. From this perspective, anger may primarily convey to an observer a readiness to attack, whereas fear may primarily convey a readiness to submit or retreat (see [54]). From this perspective behavioral intentions are considered “syndromes of correlated components” [53] (p. 151). Fridlund is not alone in these assumptions. Some researchers have gone so far as to suggest that feeling states associated with emotions are merely conscious perceptions of underlying behavioral intentions, or action tendencies, which implies that emotional feeling is simply the experience of behavioral intention, similar to William James’s theory [55].

It is worth noting that empirical research has provided support for the idea that behavioral intention is conveyed through emotional expression. For example, one study demonstrated that action tendencies and emotion labels are attributed to faces at comparable levels of consistency [55]. Similarly, in forced-choice paradigms [54], cross-cultural evidence indicates that participants assign behavioral intention descriptors with about equal consistency as they do with emotion descriptors.

A focus on approach-avoidance tendencies has been highlighted in most of the research conducted to date. The ability to detect another’s intention to approach or avoid us is thought of as a principal factor governing social exchange. However, much of the work on approach-avoidance behavioral

motivations has also tended to concentrate on the experience or response of an observer to a stimulus event [56]. One common method of operationalizing approach and avoidance then stems from traditional behavioral learning paradigms that link behavioral motivation and emotion through reward versus punishment contingencies [57]. Approach motivation is defined by appetitive, reward-related behavior, while avoidance motivation is defined by aversive, punishment-related behavior, where appetitive behavior is movement toward a reward and aversive behavior is movement away from a punishment.

Much research has focused on the relationship between approach and avoidance tendencies and emotional experience [58]. However, there has been less attention paid to whether approach and avoidance tendencies are fundamentally signaled by the external expression of emotion. It stands to reason that if the experience of emotion is associated with approach and avoidance tendencies, these tendencies should be signaled to others when expressed. This distinction is important as the approach-avoidance tendencies attributed to expressive faces may not always match the approach-avoidance reactions elicited by them. For example, the expression of joy arguably conveys a heightened likelihood of approach by the expressor and a reaction of approach from the observer [56]. In contrast, anger expressions signal approach by the expressor, but tend to elicit avoidance by the observer.

Recent insights from the embodiment literature also provide evidence that emotional experiences are grounded in specific action tendencies [34]. This means that emotional experiences can be expressed through stored action tendencies in the body, rather than through semantic cues. For example, studies have examined the coherence between emotional experience (positive or negative) and approach (arm flexion, i.e., pulling toward) versus avoidance behavior (arm extension, i.e., pushing away) [59]. In one study, participants were randomly assigned to an arm flexion (approach behavior) or arm extension (avoidance behavior) condition, either during the reading of a story about a fictional character or during a positive versus negative semantic priming task before reading the story. Participants in the congruent conditions (happy prime and arm flexion, and sad prime and arm extension) were able to remember more items from the story.

Although important for explaining behavioral responding, these studies did not address whether basic tendencies to approach or avoid were also fundamentally signaled by emotional expressions. If they were, expressions coupled with approach and avoidant behaviors should impact the efficiency of emotion recognition. In one set of studies, anger expressions were found to facilitate judgments of approach versus withdrawing faces compared with fear expressions [60]. Similarly, perceived movement of a face toward or away from an observer likewise facilitated angry or fearful expression perception [61]. Thus, approach and avoidance movement are associated in a fundamental way with the recognition of anger and fear displays, respectively, supporting the conclusion that basic action tendencies are inherently associated with the perception of emotion.

In sum, despite widely debatable assumptions about the nature of emotion and emotional expression across various

theories, most tend to agree that emotion expression conveys fundamental information regarding basic behavioral tendencies [60].

III. EMOTION: COMPUTATIONAL PRINCIPLES AND FOUNDATIONS

In this section, we aim to establish computational foundations for analyzing and recognizing emotions from visual media. Emotion recognition systems typically involve several fundamental data-related stages, including data collection (Section III-A), data reliability assessment (Section III-B), and data representation (Sections III-D, III-E, and III-F for general computer vision-based representation, movement coding, and context and function purpose detection, respectively). As we present specific examples at each stage, we will emphasize the underlying principles they adhere to. We provide a list of representative datasets in Section III-C. We will also introduce the factors of acted portrayals (Section III-G), cultural and gender dialects (Section III-H), structure (Section III-I), personality (Section III-J), and affective style (Section III-K) in inferring emotion, based on prior research.

A. Data Collection

Because the categories of emotions are not well-defined, it is not possible to program a computer to recognize all emotion categories based on a set of predefined logic rules, computational instructions, or procedures. Thus, researchers must take a data-driven approach in which computers learn from a large quantity of labeled, partially labeled, and/or unlabeled examples. To enable such research and subsequent real-world applications, it is essential to collect large-scale, high-quality, ecologically valid datasets. To highlight the complexity of the data collection problem as well as to introduce best practices, we describe a few data collection approaches that incorporate psychological principles in their design.

1) *Evoked Emotion – Immediate Response*: In the field of modeling evoked emotion, earlier researchers utilized the IAPS dataset, which consisted of only 1,082 images rated for evoked emotional response [46]. In 2017, Lu et al. [48] introduced one of the first large-scale datasets, the EmoSet, utilizing a human subject study. The EmoSet dataset is much larger, and all images are complex scenes that humans regularly encounter in daily life.

To create a diverse image collection, the researchers employed a data-crawling approach to gather nearly 44,000 images from social media, and obtained emotion labels (both dimensional and categorical) using crowdsourcing via the Amazon Mechanical Turk (AMT) platform. They used the VAD (Valence, Arousal, and Dominance) dimensional model [27], which is similar to the PAD model. The researchers followed strict psychological subject study procedures and validation approaches. The images were collected from more than 1,000 users' Web albums on Flickr using 558 emotional words as search terms. These words were summarized by Averill [62].

The researchers carefully designed their online crowdsourcing human subject study to ensure the quality of the data. For example, each image was presented to a subject for

	Valence	Arousal	Likability
High Affect			
Neutral Affect			
Low Affect			

Fig. 2. Example images in the ISEE dataset [50]. Images were selected after a thorough test-retest reliability study.

exactly six seconds. This design differed from conventional object recognition data annotation tasks, where the subject was often given no restrictions on the amount of time to view an image. This design followed psychological convention, as the intention was to collect the subject’s *immediate* affective response to the visual stimuli. If subjects were given varying amounts of time to view an image before rating it, the data would not be a reliable capture of their immediate affective response. To accommodate for this, subjects were given the option to click a “Reshow Image” button if they needed to refer back to the image. In addition, recognizing that categorical emotions may not cover all feelings, this method allowed the subject to enter other feelings they may have had.

2) *Evoked Emotion – Test-Retest Reliability*: The data collection method proposed by Lu et al. [48] aimed to understand immediate affective responses to visual content, but it did not ensure retest reliability of affective picture stimuli over time and across a population. Many psychological studies, from behavioral to neuroimaging studies, have used visual stimuli that consistently elicited specific emotions in human subject. While the IAPS and other pictorial datasets have validated their data, they have not examined the retest reliability or agreement over time of their picture stimuli.

To address this issue, Kim et al. [50] developed the Image Stimuli for Emotion Elicitation (ISEE) as the first set of stimuli for which there was an unbiased initial selection method and with images specifically selected for high retest correlation coefficients and high within-person agreement across time. The ISEE dataset used a subset of 10,696 images from the Flickr-crawled EmoSet. In the initial screening study, study

participants rated stimuli twice for emotion elicitation across a one-week interval, resulting in the selection of 1,620 images based on the number of ratings and retest reliability of each picture. Using this set of stimuli, a second phase of the study was conducted, again having participants rate images twice with a one-week interval, in which the researchers found a total of 158 unique images that elicited various levels of emotionality with both good reliability and good agreement over time. Fig. 2 shows 18 example images in the ISEE dataset.

3) *Expressed Emotion – Body*: In the field of expressed emotion recognition, the collection of data on bodily expressed emotions has received less attention compared to the more widely studied areas of facial expression and microexpression data collection. In addition, whereas earlier studies often relied on data collected in controlled laboratory environments, recent advancements in technology have made it possible to collect data in more naturalistic, real-world settings. These “in-the-wild” datasets are more challenging to collect, but they offer the opportunity to capture a more diverse range of emotions and expressions. Whereas laboratory environments provide the advantage of advanced sensors such as Motion Capture (MoCap), body temperature, and brain electroencephalogram (EEG) for collecting data, and it is possible to capture self-identified rather than perceived emotional expression, it is impossible to accurately replicate the vast array of diverse real-world scenarios within a controlled laboratory setting.

Using video clips from movies, TV shows, sporting, and wedding events as a source of data for emotion recognition has several advantages. These videos provide a wide range

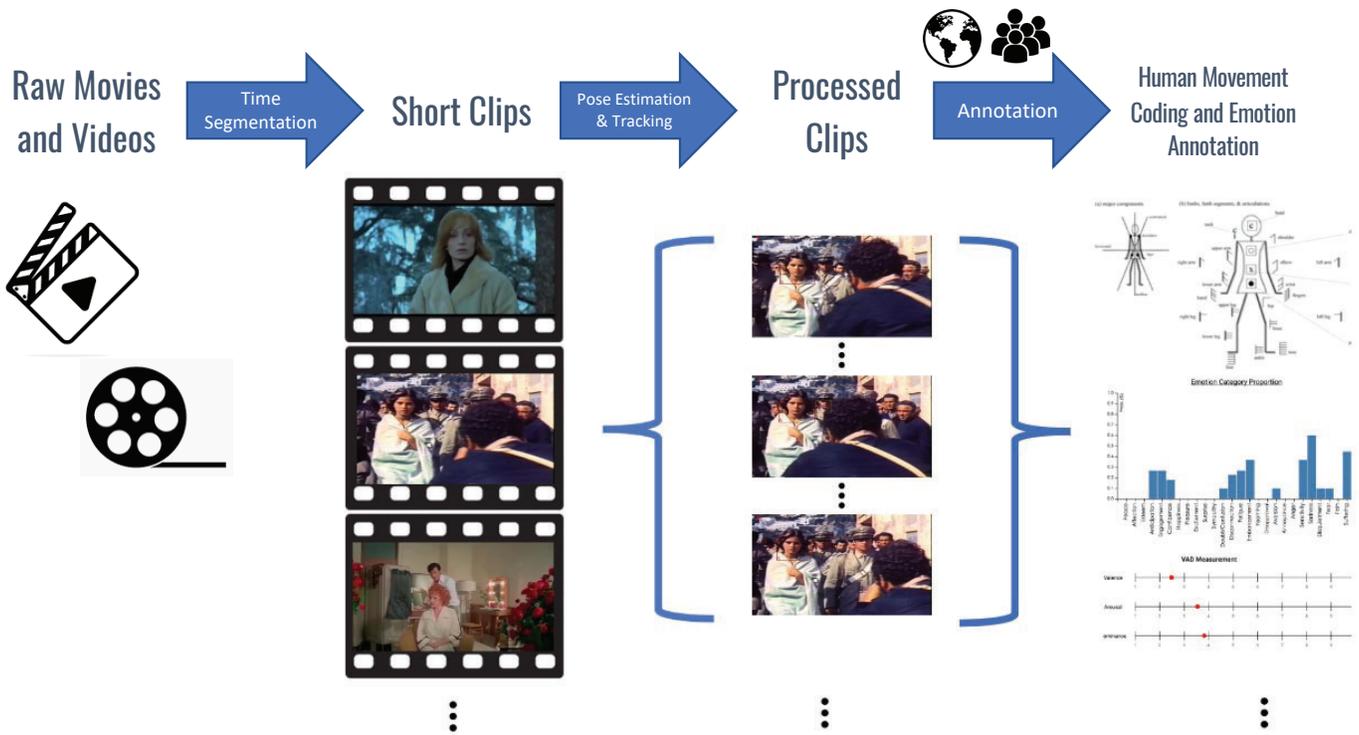


Fig. 3. A data collection pipeline was developed by Luo et al. [51] to utilize crowdsourcing to annotate vast amounts of videos available on the Internet. The pipeline involved obtaining raw movies and videos from the Internet, segmenting them into short clips, processing the clips using computer vision techniques to extract posture information and track individual characters, and utilizing crowdsourcing and expert annotation to assign emotion and movement labels to the clips.

of scenarios, environments, and situations that can be used to train computer systems to understand human behavior, expression, and movement. For instance, these videos have recorded scenes during natural and man-made disasters, providing valuable information for understanding human emotions under extreme conditions. In addition, a large proportion of video shots in movies is of outdoor human activities, providing a diverse range of contexts for training.

However, it is important to note that using publicly available video clips as a source of data has its limitations. One such limitation is that this approach can only capture perceived emotions, as opposed to self-identified emotions. In many applications, perceived emotions are a sufficient proxy for actual emotions, particularly when the goal is for robots to “perceive” or “express” emotions in a way that is similar to humans for efficient communication with humans. A further constraint is that the videos mainly feature staged or user-selected scenes, rather than depicting natural everyday interactions. This topic will be further explored in Section III-G.

Luo et al. [51] developed the first dataset for bodily expressed emotion understanding (BEEU), named the BoLD (Body Language Dataset), using this approach. The data collection pipeline is illustrated in Fig. 3. The researchers collected hundreds of movies from the Internet and cut them into short clips. An identified character with landmark tracking in a single clip is called an *instance*. They used the AMT platform for crowdsourcing emotion annotations of a total of over 48,000 instances. The emotion annotation included the VAD dimensional model [27] and 26 emotion categories [63].

B. Data Quality Assurance

Quality control is a *crucial* aspect for crowdsourcing, particularly for affect annotations. Different individuals may have varying perceptions of affect, and their understanding can be influenced by factors such as cultural background, current mood, gender, and personal experiences. Even an honest participant may provide uninformative affect annotations, leading to poor-quality data. In this case, the variance in acquiring affect usually comes from two kinds of participants, i.e., dishonest ones, who give useless annotations for economic motivation, and exotic ones, who give inconsistent annotations compared with others. The existence of exotic participants is inherent in emotion studies. The annotations provided by an exotic participant could be valuable when aggregating the final ground truth or investigating cultural or gender effects of affect. However, we typically want to reduce the risk of high variance caused by dishonest and exotic participants in order to collect generalizable annotations.

In the case of the BoLD dataset [51], five complementary mechanisms were used, including three online approaches (i.e., analyzing while collecting the data) and two offline (i.e., postcollection analysis), based on a recent technological breakthrough for crowdsourced affective data collection [49]. These mechanisms were participant emotional intelligence (EQ) screening [64], annotation sanity/consistency check [51], gold standard test based on control instances [51], and probabilistic multigraph modeling for reliability analysis [49].

Particularly critical is the probabilistic graphical model Ye

et al. developed to jointly model *subjective reliability*, which is independent from supplied questions, and *regularity* [49]. For brevity of discussion, we focus on using the mode(s) of the posterior as point estimates. We assumed each subject i had a reliability parameter $\tau_i \in [0, 1]$ and regularity parameters $\alpha_i, \beta_i > 0$ characterizing their agreement behavior with the population, for $i = 1, \dots, m$. We also used parameter γ for the rate of agreement between subjects by pure chance. Let $\Theta = (\{\tau_i, \alpha_i, \beta_i\}_{i=1}^m, \gamma)$ be the set of parameters. Let Ω_k be a random sub-sample from subjects $\{1, \dots, m\}$ who labeled the stimulus k , where $k = 1, \dots, n$. We also assumed sets Ω_k 's were created independently from each other. For each image k , every subject paired from Ω_k^2 , i.e., (i, j) with $i \neq j$, had a binary indicator $I_{i,j}^{(k)} \in \{0, 1\}$ coding whether their opinions agreed on the respective stimulus. We assumed $I_{i,j}^{(k)}$ was generated from a probabilistic process involving two latent variables. The first latent variable $T_j^{(k)}$ indicated whether subject O_j was reliable or not. Given that it was binary, a natural choice of model was the Bernoulli distribution. The second latent variable $J_i^{(k)}$, lying between 0 and 1, measured the extent to which subject O_i agreed with other reliable responses. We used the Beta distribution parameterized by α_i and β_i to model $J_i^{(k)}$ because it was a widely used and flexible parametric distribution for quantities on the interval $[0, 1]$.

In a nutshell, $T_j^{(k)}$ is a latent switch (a.k.a. gate) that controls whether $I_{i,j}^{(k)}$ can be used for the posterior inference of the latent variable $J_i^{(k)}$. Hence, the researchers referred to the model as the *Gated Latent Beta Allocation* (GLBA). A graphical illustration of the model is shown in Fig. 4. If an uninformative annotator was in the subject pool, their reliability parameter τ_i was zero, though others could still agree with their answers by chance at a rate of γ . On the other hand, if an individual was very reliable yet often provided controversial answers, their reliability τ_i could be one, while they typically disagreed with others, as indicated by their high irregularity

$$\mathbb{E}[J_i^{(k)}] = \frac{\alpha_i}{\alpha_i + \beta_i} \approx 0.$$

We were interested in finding both types of participants. Most participants were between these two extremes. The quantitative characterization of participants by GLBA will assist in selecting subsets of the data collection for quality control or gaining a comprehensive understanding of subjectivity. For more details, please refer to [49, 65].

A recent study [66] presented a Python-based software program called MuSe-Toolbox, which combines emotion annotations from multiple individuals. The software includes several existing annotation fusion methods, such as Estimator Weighted Evaluator (EWE) [67] and Generic-Canonical Time Warping (GCTW) [68]. In addition, the authors have developed a new fusion method based on EWE, named Rater Aligned Annotation Weighting (RAAW), which is also included in the software. Furthermore, MuSe-Toolbox includes the capability to convert continuous emotion annotations into categorical labels.

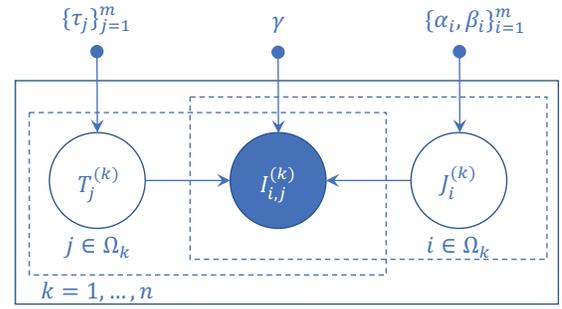


Fig. 4. Probabilistic graphical model GLBA, was developed by Ye et al. [49], to model the subjective reliability and regularity in a crowdsourced affective data collection. The work enables researchers in affective computing to effectively identify and exclude highly subjective annotation data provided by uninformative human participants, thereby improving the overall quality of the collected data.

C. Existing Datasets

Several recent literature surveys have provided an overview of existing datasets for emotions in visual media. In order to avoid duplication of effort, readers are directed to these papers for further information, which include surveys on evoked emotion [16], BEEU [17], facial expression recognition (FER) [4], microexpression recognition (MER) [7], and multimodal emotion [22, 69, 70]. Table I presents a comparison of the properties of some representative datasets. Researchers are advised to thoroughly review the data collection protocol used before utilizing a dataset to ensure that the data has been collected in accordance with appropriate psychological guidelines. In addition, when crowdsourcing is utilized, effective mechanisms are essential for filtering out uninformative annotations.

D. Data Representations

After data collection and quality assurance stages, a significant technological challenge is to represent the *emotion-relevant information* present in the raw data in a concise form. Whereas current deep neural network (DNN) approaches often utilize raw data, such as matrices of pixels, as input in the modeling process, utilizing a compact data representation can potentially improve the efficiency of the learning process, allowing for larger-scale experiments to be conducted with limited computational resources. In addition, a semantically meaningful data representation can facilitate interpretability, which is crucial for certain applications. There are numerous methods for compactly representing raw visual data, and we discuss several intriguing or widely used data representations for emotion modeling in the following.

1) Roundness, Angularity, Simplicity, and Complexity:

Colors and textures are commonly used in image analysis tasks to represent the content of an image, but research has shown that shape can also be an effective representation when analyzing evoked emotions. In both visual art and psychology, the characteristics of shapes, such as *roundness*, *angularity*, *simplicity*, and *complexity*, have been linked to specific emotional responses in humans. For example, round and simple shapes tend to evoke positive emotions, while

TABLE I
RECENT REPRESENTATIVE DATASETS FOR EMOTION RECOGNITION

Dataset Name	Labeled Samples	Data Type	Categorical Emotions	Continuous Emotions	Lab-Controlled	Year	Primary Application
IAPS [45]	1.2k	I	-	VAD		2005	Evoked
FI [71]	23.3k	I	8 [‡]	-		2016	Evoked
VideoEmotion-8 [72]	1.2k	V	8 [†]	-		2014	Evoked
Ekman-6 [73]	1.6k	V	6 [†]	-		2018	Evoked
E-Walk [74]	1k	V	4 [‡]	-		2019	BEEU
BoLD [51]	13k	V	26 [†]	VAD		2020	BEEU
iMiGUE [75]	0.4k	V	2	-		2021	BEEU*
CK+ [76]	0.6k	V	7 [‡]	-	✓	2010	FER
Aff-Wild [77]	0.3k	V	-	VA		2017	FER
AffectNet [78]	450k	I	7 [‡]	-		2017	FER
EMOTIC [79]	34k	I	26 [†]	VAD		2017	FER*
AFEW 8.0 [80]	1.8k	V	7 [‡]	-		2018	FER
CAER [81]	13k	V	7 [‡]	-		2019	FER*
DFEW [82]	16k	V	7 [‡]	-		2020	FER
FERV39k [83]	39k	V	7 [‡]	-		2022	FER
SAMM [84]	0.2k	V	7 [‡]	-	✓	2016	MER
CAS(ME) ² [85]	0.06k	V	4 [‡]	-	✓	2017	MER
ICT-MMMO [86]	0.4k	V,A,T	-	Sentiment		2013	Multi-Modal
MOSEI [87]	23.5k	V,A,T	6 [†]	Sentiment		2018	Multi-Modal

[†] A superset of Ekman's basic emotions [‡] Ekman's basic emotions + neutral [§] Mikels' emotions

* Micro-gesture understanding and emotion analysis dataset * Context-aware emotion dataset
Data Type Key: (I)mage, (V)ideo, (A)udio, (T)ext

angular and complex shapes evoke negative emotions. Leveraging this understanding, Lu et al. developed a system that predicted evoked emotion based on line segments, curves, and angles extracted from an image [46]. They used ellipse fitting to implicitly estimate roundness and angularity and used features from the visual elements to estimate complexity. Later, they developed algorithms to explicitly estimate these representations [48]. Fig. 5 shows some example images with different levels of roundness, angularity, and simplicity. The researchers found that these three physically interpretable visual constructs achieved comparable classification accuracy to the hundreds of shape, texture, composition, and facial feature characteristics previously examined. This result was thought-provoking because just a few numerical-value representations could effectively predict evoked emotions.

2) *Facial Action Coding System (FACS) and Facial Landmarks*: People use particular facial muscles to express certain facial expressions. For instance, people can express anger by frowning and pursing their lips. Consequently, each facial expression can be viewed as a combination of some facial muscle movements. Ekman and Friesen developed the FACS in 1976, which encodes all movements of facial muscles [88]. FACS defines a total of 32 atomic facial muscle actions, called Action Units (AUs), including Lids Tight (AU7), Cheek Raise (AU6), and so on. By detecting all AUs of a person and linking them to specific expressions, we can identify the individual's facial expressions.

The problem of AU detection can be approached as a multilabel binary classification problem for each AU. Early work on AU detection used facial landmarks to identify regions of corresponding muscles and then applied neural networks [89] or Support Vector Machines (SVMs) [90] for classification. More recent work has developed end-to-end

AU detection networks [91]. Survey papers provide detailed introductions to facial AU detection [92, 93] and face landmark detection [94, 95]. Some researchers also used facial landmarks directly as a representation of facial information in their recognition work.

3) *Body Pose and Body Mesh*: People can express emotions through body posture and movement. By manipulating the positioning of body parts (e.g., the shoulders, arms), people produce various postures and movements. The coordinates of human joints can serve as a representation of body language, reflecting the individual's bodily expression. In the field of computer vision, two-dimensional (2D) pose estimation is a well-studied task for detecting the 2-D position of human joints in an image. Leveraging large-scale 2-D pose datasets (e.g., COCO [96]), researchers have proposed several high-performing pose networks [97, 98]. Even with challenging scenes, such as crowded or occluded scenes, these networks are able to provide comprehensive joint detection and linking.

However, 2-D pose estimation does not fully capture the three-dimensional (3D) nature of human posture and movement. 3-D human pose estimation, on the other hand, aims to predict the 3-D coordinates of human joints in space. Single-person 3-D pose estimation methods determine the 3-D joint coordinates relative to the person's root joint (i.e., the torso) [99, 100]. In addition, some multiperson 3-D pose estimation approaches comprehensively estimate the absolute distance between the camera and the individuals in the image [101].

3D human mesh estimation, which provides the 3-D coordinates of each point on the human mesh, is a further extension of the 3-D pose estimation. Researchers often utilize SMPL [103, 104] or other human graph models to represent the mesh. Fig. 6 illustrates an example of 2-D pose, 3-D

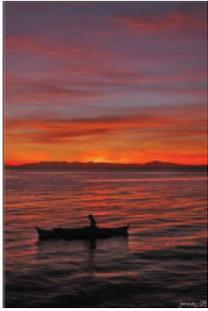
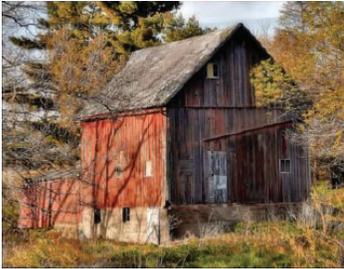
	Roundness	Angularity	Simplicity
High			
Medium			
Low			

Fig. 5. Researchers developed algorithms to compute the roundness, angularity, and simplicity/complexity of a scene as the representation for estimating the evoked emotion [48].

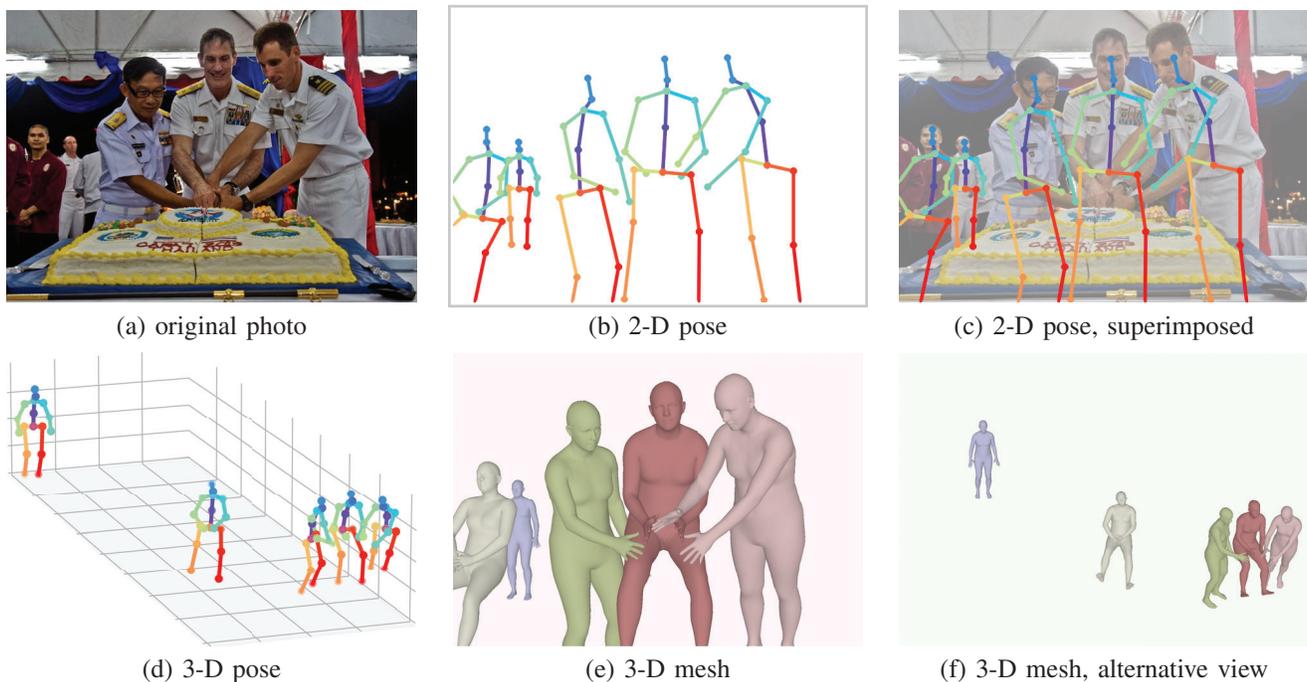


Fig. 6. An example of state-of-the-art deep learning-based human pose and mesh detection with individuals in the scene separated properly [102]. The input to the methods is a standard 2-D RGB photograph. The 3-D mesh method provides an estimation of the relative depths of the people in the scene.

pose, and 3-D mesh automatically generated from a scene with multiple people [102].

While human pose and human mesh representations provide a higher level of abstraction compared to low-level representations such as raw video or optical-flow, there is still a significant gap between these intermediate-level representations and high-level emotion representations. To bridge this gap, an intermediate-level representation that effectively describes human movements is proposed. Specifically, Laban Movement Analysis (LMA) is a movement coding system initially developed by the dance community, similar to sheet music for describing music. Fig. 7 illustrates the layers of data representation for BEEU, from low pixel-level representations to the ultimate high-level emotion representation. Elevating each layer higher in this information pyramid requires considerable advancements in abstraction technology.

E. Human Movement Coding and LMA

Expressive human movement is a complex phenomenon. The human body has 360 joints, many of which can move various distances at different velocities and accelerations, and in two (and depending on the joint, sometimes more) different directions, resulting in an astronomical number of possible combinations. These variables create an infinite number of movements and postures that can convey different emotions. Beyond this array of body parts moving in space, expressive movement also involves multiple qualitative motor components that appear in unique patterns in different individuals and situations. The complexity of human movement thus raises the question: *How do we determine which of the numerous components present in expressive movement are significant to code for emotional expression in movement?* Thus, when choosing a coding system, the early stages of each research project can benefit from deeply considering which aspects of movements are central to the expression being studied. A multistage methodology, such as first identifying what is potentially relevant and then using preliminary analyses to refine the selection of movements most promising to code, can be helpful before selecting a method to code or quantify the multitude of variables present in unscripted movement (e.g., [105, 106]).

After deliberating about which movement variables are relevant and meaningful, we must then consider the three main types of coding systems that have been used in various fields such as psychology, computer vision, animation, robotics, and AI, specifically:

- Lists of specific motor behaviors that have been found in scientific studies to be typical to the expressions of specific emotions, such as head down and moving slowly as characterizing sadness; moving backward and bringing the arms in front of the body as characterizing fear; jumping, expanding and upward movements as characterizing happiness, and so on. (for review of these studies and lists of these behaviors see [107, 108]).
- Kinematic description of the human body models, such as skeleton-based models, contour-based models, or volume-based models. Most work in the field of emotion recognition is based on skeleton-based models [109]. This

type of model uses 3-D coordinates of markers that were placed on (using various MoCap systems) or were mapped (using pose estimation techniques) to the main joints of the body, to create a moving “skeleton,” which enables researchers to quantitatively capture the movement kinematics (e.g., [110, 111]).

- LMA, a comprehensive movement analysis system that depicts qualitative aspects of movement and, theoretically [112, 113], as well as through scientific research [114, 115], relates the different LMA motor elements (movement qualities) to cognitive and emotional aspects of the moving individual.

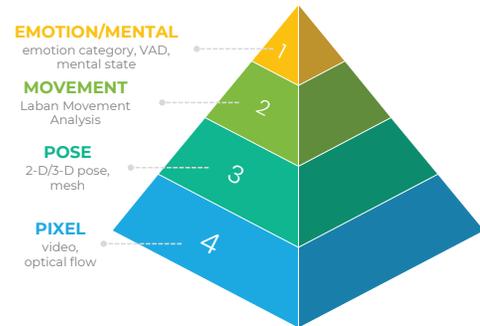


Fig. 7. For BEEU, Laban Movement Analysis as a movement coding system can serve as an intermediate level of representation to bridge the significant gap between the pose and the emotion/mental levels [51].

The first coding system, based on lists of motor behaviors, was primarily used in earlier studies in the field of psychology, where the encoding and decoding of motor behaviors into emotions and vice versa were done manually by human coders in a labor-intensive process. Other limitations of this coding system include the following:

- It is based on a limited number of behaviors that have been used in prior scientific studies. However, people can physically express emotions in many different ways, so the list of previously observed and studied whole-body expressions may not be exhaustive or inclusive of cultural variations.
- Likewise, because each study used different lists of behaviors, this method makes it difficult to compare results or to review them additively to arrive at larger verification. Thus, this coding system may miss parts of the range of bodily emotional expression such as those never observed and coded before. This limitation is especially pronounced because many of these previous studies examined emotional expressions performed by actors, whose movement tended to rely upon more stereotypical bodily emotional expressions that were widely recognized by the audience, rather than naturally occurring motor expressions.

When using the second coding system, kinematic description, in particular the skeleton-based models, researchers usually employ a set of markers similar to, or smaller than that provided by the Kinect Software Development Kit, and transform the large amount of 3-D data into various low- and/or high-level kinematic and dynamic motion features, which

are easier to interpret than the raw data. Researchers using this method have studied specific features such as head and elbow angles [111], maximum acceleration of hand and elbow relative to spine [116], distance between joints [117], among others. For a review of such studies, please refer to [17]. In recent years, instead of computing handcrafted features on the sequence of body joints, researchers have employed various deep learning methods to generate representations of the dynamics of the human body embedded within the joint sequence, such as spatiotemporal graph convolutional network (ST-GCN) [51, 118, 119]. Although the 3-D data from joint markers can provide a relatively detailed, objective description of whole-body movement, this coding system has two main limitations:

- Movement is often captured by a camera from a single view (usually the frontal), which can result in long segments of missing data from markers that are hidden by other body parts, people, or objects in a video frame. Automatic imputations of such missing data are often impractical as they tend to create incorrect and unrealistic movement trajectories.
- The SDK system has only three markers along the torso, which are insufficient for capturing subtle movements in the chest area, movements that are usually observed during authentic emotional expressions, as opposed to acted (and often exaggerated) bodily emotional expressions. Another disadvantage is that these systems have not yet been able to successfully and reliably detect many qualitative changes in movement that are significant for perceiving emotional expression.

In contrast to the quantitative data from joint markers, which enable the capture of detailed movement of every body part, the third coding system mentioned above, LMA, describes *qualitative* aspects of movement and can relate to a general impression from movements of the entire body or to the movement of specific body parts. By identifying the association between LMA motor components and emotions, and characterizing the typical kinematics of these components using high-level features (e.g., [120, 121, 122, 51, 123, 124]), researchers can overcome the limitations of other coding systems. If people express their emotions with movements that have never been observed in previous studies, we can still decode their emotions based on the quality of their movement. Similarly, if parts of the body that are usually used to express a certain emotion are not visible, it is possible that the emotion could still be decoded by identifying the motor elements composing the visible part of the movement. Moreover, by slightly changing the kinematics of a movement of a robot or animation (i.e., adding to a gesture or a functional movement the kinematics of certain LMA motor elements associated with a specific emotion), we can “color” this functional movement with an expression of that emotion, even when the movement is not the typical movement for expressing that emotion (e.g., [125, 126]). Similarly, identifying the quality of a movement can enable decoding the expressed emotion even from functional actions such as walking [127] or reaching and grasping. These advantages and the fact that

LMA features have been found to be positively correlated with emotional expressions are why LMA coding is becoming popular in studies that encode or decode bodily emotion expressions (e.g., [51, 123]). In addition, LMA offers the option to link our coding systems to diverse ways in which humans talk about and describe expressive movement—it is a comprehensive movement-theory system that is and can be used across disciplines for application in acting [128], therapy [129], education [130], and animation [131], among others. The last advantage to consider is that LMA is a comprehensive theory of body movement, much like art theory or music theory, including theories of harmony, and thus has been used by artists to attune to aesthetics including movement-perception of visual art (such as that discussed in Section VI-A) and visual, auditory and movement elements of film and theatre (discussed in Section III-G). Like music theory, LMA is capable of attending to rhythm and phrasing as elements shift and unfold over time, aspects that may be crucial to communicating and interpreting emotion expression.

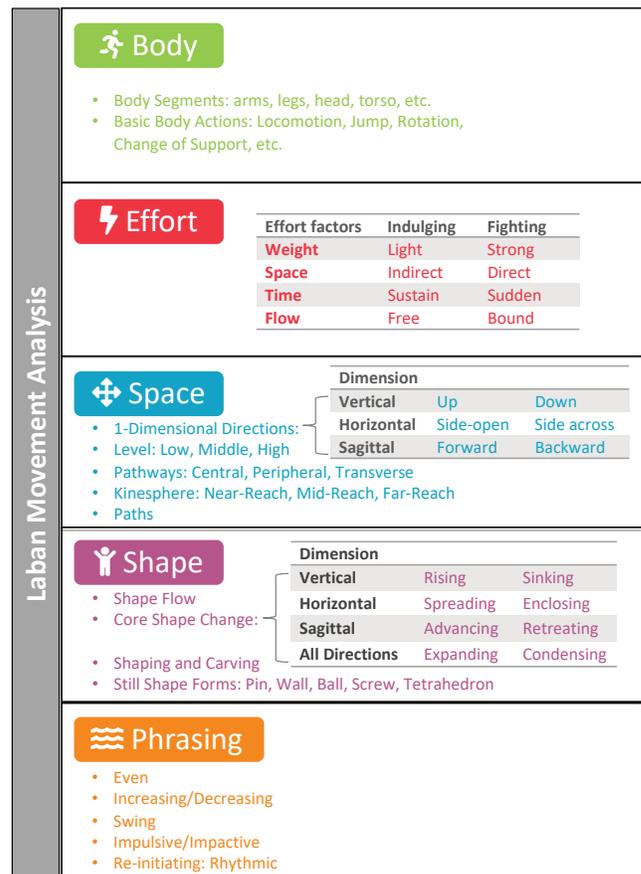


Fig. 8. Some basic components of LMA that are often used in coding.

LMA identifies four major categories of movement: **Body**, **Effort**, **Shape**, and **Space**. Each category encompasses several subsets of motor components (LMA terms are spelled with capital letters to differentiate them from regular usage of these words). Fig. 8 illustrates some basic components of LMA that are often used in coding.

The **Body** category describes what is moving and it is

composed of the elements of *Body Segments* (e.g., arms, legs, head), their coordination, and basic *Body Actions* like Locomotion, Jump, Rotation, Change of Support, and so on.

The **Effort** category describes qualitative aspect of movement, or how we move. It expresses a person's inner attitude toward movement, and it has four main factors, each describing the continuum between two extremes: indulging in the motor quality of that factor and fighting against that quality. Effort has four factors:

- *Weight Effort*, meaning the amount of force or pressure exerted by the body. Activated Weight Effort can be Strong or Light. Alternatively, there may be a lack of weight activation when we give in to the pull of gravity, which we describe as Passive or Heavy Weight.
- *Space Effort* describes attention to space, denoting the focus or attitude toward a chosen pathway, i.e., is the movement Direct or flexibly Indirect.
- *Time Effort*, describing the mover's degree of urgency or acceleration/deceleration involved in a movement, i.e., is the movement Sudden or Sustained.
- *Flow Effort*, reflecting the element of control or the degree to which a movement is Bound, i.e., restrained or controlled by muscle contraction (usually cocontraction of agonist and antagonist muscles), versus Free, i.e., being released and liberated.

The **Shape** category reflects why we move: Shape describes how the body adapts its shape as we respond to our needs or the environment: Do I want to connect with or avoid something, dominate or cower under? The way the body sculpts itself in space reflects a relationship to self, others, or to the environment. This component includes *Shape Flow* which describes how the body changes to relate to oneself; it includes *Shape Change* which describes changes in the form or shape of the body; and includes the motor components of Expanding the body or Condensing it in all directions and Rising or Sinking in the vertical dimension, Spreading and Enclosing in the horizontal dimension and Advancing or Retreating in the sagittal dimension. Another Shape component is *Shaping and Carving* which describes how a person shapes their body to shape or affect the environment or other people. For example, when we hug somebody, we might shape and carve the shape of our body, adjusting it to the shape of the other person's body, or we might shape our ideas by carving or manipulating them through posture and gesture.

The **Space** category describes where the movement goes in the environment. It describes many spatial factors such as the *Direction* where the movement goes in space, such as: Up and Down in the vertical dimension, Side open and Side across in the horizontal dimension, and Forward and Backward in the sagittal dimension; the *Level* of the movement in space relative to the entire body or parts of the body, such as Low level (movement toward the ground), Middle level, (movement maintaining level, without lowering or elevating) or High level (moving upward in space); *Paths* or how we travel through space by locomoting; and Pathways through the *Kinesphere* (the personal bubble of reach-space around each mover that can be moved in and through without locomoting or traveling.)

Movement in the Kinesphere might take Central pathways, crossing the space close to the mover's body center, Peripheral pathways along the periphery of the mover's reach space, or Transverse pathways cutting across the reach space.

In addition, another important aspect of LMA which is particularly helpful and meaningful to expression is the **Phrasing** of movements. Phrasing describes changes over time, such as changes in the intensity of the movement over time, similar to musical phrases, which can be Increasing, Decreasing, or Rhythmic, among others. It can also depict how movement components shift during the same action or a series of actions occurring over time, for example beginning emphatically with strength and then ending by making a light direct point conclusion.

Previous research has highlighted the lack of a notation system that directly encodes the correspondence between bodily expression and body movements in a way similar to FACS for face [51, 105]. LMA, by its nature, has the potential to serve as such a coding system for emotional expressions through body movement. Shafir et al. [115] identified the LMA motor components (qualities), whose existence in a movement could evoke each of the four basic emotions: anger, fear, sadness, and happiness. Melzer et al. [114] identified the LMA components whose existence in a movement caused that movement to be identified as expressing one of those four basic emotions. In an additional experiment by the Shafir group, Gilor, for her Master's thesis, has been studying the motor elements that are used for expressing sadness and happiness. In this series of studies, the LMA motor components found to evoke these emotions through movement were, for the most part, the same as those used to identify or recognize each emotion from movement, or to express each emotion through movement. For example, anger was associated with Strong Weight Effort, Sudden Time Effort, Direct Space Effort, and Advancing Shape during both emotion elicitation and emotion recognition. Fear was associated with Retreating Shape, moving backward in Space for both emotion elicitation and recognition. In addition, Enclosing and Condensing Shape, and Bound Flow Effort, were also found for emotion elicitation through movement. Sadness was associated with Passive Weight Effort, Sinking Shape, and Head drop for emotion elicitation, emotional expression, and emotion recognition, and arms touching the upper body was also a significant indicator for emotion elicitation and expression. Sadness expression was also associated with using Near-Reach Space and Stillness. In contrast, Happiness was associated with jumping, rhythmic movement, Free and Light Efforts, Spreading and Rising Shape, and moving upward in Space for emotion elicitation, emotion expression, and emotion recognition. Happiness expression was also associated with Sudden Time Effort and Rotation. These findings for Happiness were also validated by van Geest et al. [132]. Whereas these studies represent a promising start, further research is needed to create a comprehensive LMA coding system for bodily emotional expressions.

F. Context and Functional Purpose

While human emotions are shown through the face and body, they are closely connected to context and purpose. Thus,

the image context surrounding the human can also be used to further identify human emotions. The context information includes the semantic information of the background and what the person is holding, which can assist in identifying activities of the human, thereby allowing for more accurate prediction of the human's emotion. For instance, people are more likely to feel happy than sad during a birthday party. Context information also includes interactions among humans, which can help infer emotion. For example, people are more likely to be angry when they are engaged in a heated argument with others.

Since the early 2000s, researchers in the field of image retrieval have been developing context recognition systems using machine learning and statistical modeling [133, 134, 135]. With the advent of deep learning and the widespread use of modern graphics processing units (GPUs) or AI processors, accurate object annotation from images has become more feasible. Several deep learning-based approaches have been proposed to leverage context information to enhance basic emotion recognition networks [136, 81, 137]. Sections IV-B3 and IV-E provide more details.

In addition to contextual information, the functional purpose of a person's movement can also provide valuable insights when inferring emotions. Movement is a combination of both functional and emotional expression, and thus, emotion recognition systems must be able to differentiate between movement that serves a functional purpose and movement that expresses emotions. Action recognition [138], an actively studied topic in computer vision, has the potential to provide information on the function of a person's movement and assist in disentangling functional and emotional components of movement.

G. Acted Portrayals of Emotion

Research on emotions has often turned to acted portrayals of emotion to overcome some of the challenges inherent in accessing adequate datasets, particularly because evoking intense, authentic emotions in a laboratory can be problematic both ethically [139] and practically. This is because, as noted in Section II, emotional responses vary. Datasets relying upon actors have been useful in overcoming challenges of obtaining ample, adequately labeled emotion-expression data because compared to unscripted behavior ("in the wild"), where emotion expression often appears in blended or complex ways, actors are typically serving a narrative, in which emotion is part of the story. In addition, sampling emotional expression in the wild encounters cultural distinctions for the expressions themselves, as well as social norms for emotion expression or its repression [140], which may also be culturally scripted for gender. Thus, researchers interested in emotional expressivity and nonverbal communication of emotion often turn to trained actors [139] both to generate new datasets of emotionally expressive movement (e.g., [141]) and for the sampling of emotion expression (e.g., [51]). Such datasets are useful because actors coordinate all three channels of emotions expression, namely vocal, facial, and body, to produce natural or authentic-seeming expressions of emotions. Some

researchers have validated the perception of actor-based and expert movement-based datasets in the lab by showing them to lay observers [141, 114]. This approach also entails problems, in that whereas it may capture norms of emotion expression and its clear communication, it may miss distinctions related to demographics such as gender [142], ethnolinguistic heritage, individual, or generational norms [143, 144]. According to cultural dialect theory, we all have non-verbal emotion accents [144], meaning emotion is expressed differently by different people in different regions, from different cultures. Only some of those cultural dialects appear when sampling films. Such films have often been edited so that viewers beyond that cultural dialect can "read" the emotional expression central to the narrative. Nuance in nonverbal dialects may be excluded in favor of general appeal.

Yet, an advantage of generating datasets from actors or other trained expressive movers is that ground truth can be better established. The intention of the emotion expressed, the emotion felt, and later the emotion perceived from viewing can all be assessed when generating the dataset. Likewise, because actors coordinate image, voice, and movement in the service of storytelling, the context and purpose are clarified, and thus many of the multiple expressive modes can be organized into individual emotion expression [145, 146]. Moreover, because performing arts productions, such as movies and films made of theatre, music, and dance performances, integrate multiple communication modes, the creative team collaborates frequently about the emotional tone or intent of each work, coordinating lighting, scenery/background, objects, camerawork, and sound, with the performers (actors, dancers, musicians). While the team articulates their intentions during the creative process, the resulting produced art often resonates with wider variation to different audiences, according to their perceptions and tastes.

For researchers relying upon acted emotions, it may be helpful to understand some ways in which actor training considers the role of emotions in theatre and film [147]. The role of emotion in narrative arts may reflect some of the theories about the role of emotion itself [148]—to approximately inform the character (organism) about their needs, in order to drive them to take action to meet their needs, or to provide feedback on how recent actions meet or do not meet their needs. As actors prepare, they identify a character's needs (called their objective) and are moved by emotion to drive the character's action to overcome obstacles as they pursue the character's objective. Thus, when collecting data from acted examples, emotion expression can often be found preceding action in the narrative, or in response to it. Actors are also trained to listen with their whole being to their scene partner, to "hold space" during dialogue for emotion expression to fully unfold and complete its purpose of moving either the speaker or the listener to response or action. This expectation is particularly true in opera or other musical theatre genres, which often extend the time for emotion expression.

In terms of 3-D modeling, actors trained for theatre not only highly develop the specific emotion-related action tendencies of the body but also consider viewer perception from 3D, for example, when performing on a thrust stage or theatre-

in-the-round. Thus, their emotion expression may be more easily picked up by systems marking parts of the body during movement. Because bodily emotion expression is so crucial to a narrative, an important application of this field might be to automate audio description of emotion expression in film for the visually impaired or audio description of movement features salient to emotion expression.

Whereas we often do not recognize subtle emotion expression in strangers as well as we can in those we know, when actors portray characters within the circumstances of a play, the audience, gradually over the arc of the play, comes to perceive the emotional expression of each character as revealed over time. Understanding how this works in art can help us develop systems that take the time to learn and better understand emotion expression in diverse contexts and individuals, similar to how current voice recognition learns individual accents.

H. Cultural and Gender Dialects

In a meta-analysis examining cultural dialects in the non-verbal expression of emotion, measurable variations were found in response to facial expression across cultures [149]. Moreover, it has long since been acknowledged that learned display rules [150] and decoding rules [151] vary across different cultures. For example, in Japanese culture, overt displays of negative emotional expression are discouraged. Further, smiling behavior in this context is often considered an attempt to conceal negative emotions. In contrast, in American culture, negative expressions are considered more appropriate to display [151].

Previous research has similarly demonstrated a powerful influence of gender-related emotion stereotypes in driving differential expectations for the type of emotions men and women are likely to experience and express. Men are expected to experience and express more power-related emotions like anger and contempt, whereas women are expected to experience and express more powerless emotions like sadness and fear [152]. These findings match self-reported emotional experiences and expressions with strong cultural consistency across 37 different countries worldwide [153]. Furthermore, there are even differences in the extent to which neutral facial appearance resembles different emotions, with male faces physically resembling anger expressions more (low, hooded brows, thin lips, angular features) and female faces resembling fear expressions more [154].

While cultural norms affect how and whether emotions are displayed, such norms also influence how displays of emotion are perceived. For example, when a culture's norm is to express emotion, less intensity is read into such displays, whereas, in cultures where the norm is not to express emotion intensely, the same displays are read as more intense [155]. In this way, visual percepts derived from objective stimulus characteristics can generate different subjective experiences based on culture and other individual factors. For instance, there is notable cultural variation in the extent to which basic visual information, such as background versus foreground, is integrated into observers' perceptual experiences [156].

Culture and gender add complexity to human emotional expression, yet little research to date has examined individual variation in responses to visual scenes, either in terms of basic aesthetics or the emotional responses people have. Future work assessing simple demographic details (e.g., gender, age) will begin to explore this important source of variation.

I. Structure

Some basic visual properties have been found that characterize positive versus negative experiences and preferences. Most notably, round features—whether represented in faces or objects—elicit feelings of positivity and warmth and tend to be preferred over angular visual features. This preference has been used to explain the roundness of smiling faces, and the angularity of anger displays [157]. Such visual curvature has also been found to influence attitudes toward, and preference for even meaningless patterns represented in abstract visual designs [158]. The connection between affective responses and these basic visual forms has helped computer vision predict emotions evoked from pictorial scenes, as mentioned earlier [46].

Importantly, the dimensional approach to assessing visual properties underlying emotional experience can be used to examine both visual scenes and faces found within those scenes. Indeed, the dimensional approach adequately captures both “pure” and mixed emotional facial expressions [159], as well as affective responses to visual scenes, as demonstrated by the IAPS. Critically, even neutral displays have been found to elicit strong spontaneous impressions [160], ones that are effortless, nonreflective, and highly consensual. Recent research utilizing computer-based models suggests that these inferences are largely driven by the same physical properties found in emotional expressions. For instance, Said and colleagues [161] employed a Bayesian network trained to detect expressions in faces, and then applied this analysis to images of neutral faces that had been rated on a number of common personality traits. The results showed that the trait ratings of faces were meaningfully associated with the perceptual resemblances that these “neutral” faces had with emotional expressions. Thus, these results speak to a mechanism of perceptual overlap whereby expression and identity cues can both trigger similar emotional responses.

A reverse engineering model of emotional inferences has suggested that perceptions of stable personality traits can be derived from emotional expressions as well [162]. This work implicates appraisal theory as a primary mechanism by which observing facial expressions can inform stable personality inferences made of others. This account suggests that people use appraisals that are associated with specific emotions to reconstruct inferences of others' underlying motives, intents, and personal dispositions, which they then use to derive stable impressions. It has likewise been shown that emotion-resembling features in an otherwise neutral face can drive person perception [152]. Finally, research has also suggested that facial expressions actually take on the form that they do to resemble static facial appearance cues associated with certain functional affordance, such as facial maturity [163], and gender-related appearance [152].

J. Personality

Personality describes an individual’s relatively stable patterns of thinking, feeling, and behaving. These are characterized by certain personality traits which represent a person’s disposition towards the world. Emotions, on the other hand, are the consequence of the individual’s response to external or internal stimuli. They may change when the stimuli change, and therefore are considered as states (as opposed to traits). Just as a full-blown emotion represents an integration of feeling, action, appraisal, and wants at a particular time and location, personality represents integration of these components over time and space. Researchers have tried to characterize the relationships between personality and emotions. Several studies found correlations between certain personality traits and specific emotions. For instance, the trait neuroticism was found to be correlated with negative emotions, while extroverted people were found to experience higher levels of positive emotions than introverted people [164]. These correlations were explained by relating different personality traits to specific emotion regulation strategies e.g., [165] or by demonstrating that evaluation mediates between certain personality traits and negative or positive affect [166]. Whatever the reason for these correlations is, the fact that they are correlated might help to create state-to-trait inferences. Therefore, if emotional states can be mapped to personality, the ability to automatically recognize emotions could provide tools for automatic detection of personality.



Fig. 9. The Big Five personality traits, or OCEAN.

Advances in computationally, data-driven methods offer promising strides toward personality traits being predicted based on a dataset of individuals’ behavior, self-reported personality traits, or physiological measures. It is also possible to use factor analyses to identify underlying dimensions of personality, such as the Big Five personality traits, or OCEAN (openness, conscientiousness, extraversion, agreeableness, and

neuroticism) [167, 168] (Fig. 9). Personality can likewise be inferred from facial expressions of emotion [162] and even from emotion-resembling facial appearance [169]. Machine learning algorithms have recently been successfully employed to predict human perceptions of personality traits based on such facial emotion cues [170]. Further, network analysis, such as social network analysis, can also be incorporated to identify patterns of connectivity between different personality traits or behavioral measures. Finally, interpreting a person’s emotional expression might also depend on the perception of their personality, when additional data is available for inferring personality [171]. For further information, readers can refer to recent research in this area, e.g., [172, 173].

K. Affective Style

Affective styles driven by the tendencies to approach reward versus avoid punishments found their way into early conflict theories [174] and remain a mainstay of contemporary psychology in theories such as Carver and Scheier’s Control Theory [175] and Higgins’s Self-discrepancy Theory [176].

Evidence for the specific association between emotion and approach-avoidance motivation has largely involved the examination of differential hemispheric asymmetries in cortical activation. Greater right frontal activation has been associated with avoidance motivation as well as with flattened positive affect, and increased negative affect. Greater left frontal activation has been associated with approach motivation and positive affect [177]. Supporting the meaningfulness of these findings, Davidson argued that projections from the mesolimbic reward system, including basal ganglia and ventral striatum, which are associated with dopamine release, give rise to greater left frontal activation. Projections from the amygdala associated with the release of the primary vigilance-related transmitter norepinephrine give rise to greater right frontal activation.

Further evidence supporting the emotion/behavior orientation link stems from evidence accumulated in studies using measures of behavioral motivation based on Gray’s [178] proposed emotion systems. The most widely studied of these are the Behavioral Activation System (BAS) and the Behavioral Inhibition System (BIS). The BAS is argued to be highly related to appetitive or approach-oriented behavior in response to reward, whereas the BIS is argued to be related to inhibited or avoidance-oriented behavior in response to punishment. Carver and White [179] developed a BIS/BAS self-report rating measure that is thought to tap into these fundamental behavioral dispositions. They found that extreme scores on BIS/BAS scales were linked to behavioral sensitivity toward reward versus punishment contingencies, respectively [179]. BIS/BAS measures have been shown to be related to emotional predisposition, with positive emotionality being related to the dominance of BAS over BIS and depressiveness and fearful anxiety being related to the dominance of BIS over BAS [180].

Notably, for many years there existed a valence (positive/negative) versus motivational (approach/avoidance) confound in all work conducted in the emotion/behavior domain. Negative emotions were associated with avoidance-oriented and positive emotions with approach-oriented behavior, a

contention supported by much of the work reviewed above. The valence/motivation confound led researcher Harmon-Jones and colleagues to test for hemispheric asymmetries in activation associated with anger, a negative emotion with an approach orientation (aggression) [181]. They argued that if left hemispheric lateralization was associated with anger this would indicate that the hemispheric lateralization in activation previously found was in fact due to behavioral motivation. However, right hemispheric lateralization would indicate that they were due to valence. In these studies, they found that dispositional anger [181] was associated with left lateralized EEG activation, consistent with the first interpretation and with that previously reported only for positive emotion. They supported this conclusion by showing that the dominance of BAS over BIS was associated with anger [58].

IV. EMOTION RECOGNITION: KEY IDEAS AND SYSTEMS

The field of computer-based emotion recognition from visual media is nascent, but has seen encouraging developments in the past two decades. Interest in this area has also grown sharply (Section IV-A). We will highlight some existing research ideas and systems related to modeling evoked emotion (Section IV-B), facial expression (Section IV-C), and bodily expression (Section IV-D). In addition, we will discuss integrated approaches to emotion recognition (Sections IV-E and IV-F). Because of the breadth of the field, it is not possible to cover all exciting developments, so we will focus our review on the evolution of the field and some of the most current, cutting-edge results.

A. Exponential Growth of the Field

To gain insight into the growing interest of the IEEE and computing communities in emotion recognition research, we conducted a survey of publications in the IEEE Xplore and ACM Digital Library (DL) databases. Results revealed an exponential increase in the number of publications related to emotion or sentiment in images and videos over the last two decades (Fig. 10).

As of February 2023, a total of 48,154 publications were found in IEEE Xplore, with the majority (35,247 or 73.2%) being in conference proceedings, followed by journals (8,214 or 17.1%), books (2,400 or 5.0%), magazines (1,413 or 2.9%), and early access journal articles (831 or 1.7%). The field has experienced substantial growth, with a 25-fold increase during the period from the early 2000s to 2022, rising from an average of 275 publications per year to about 7,000 per year in 2022.

In the ACM DL, a total of 30,185 publications were found, with conference proceedings making up the majority (24,817 or 82.2%), followed by journals (3,851 or 12.8%), magazines (780 or 2.6%), newsletters (505 or 1.7%), and books (264 or 0.9%). In the ACM community, the field has seen a 22-fold growth during the same period, with an average of 170 publications per year in the early 2000s rising to about 3,800 per year in 2022.

A baseline search indicated that the field related to images and videos had a roughly linear, 6-fold growth during the same period. Emotion-related research accounted for 14.2% of

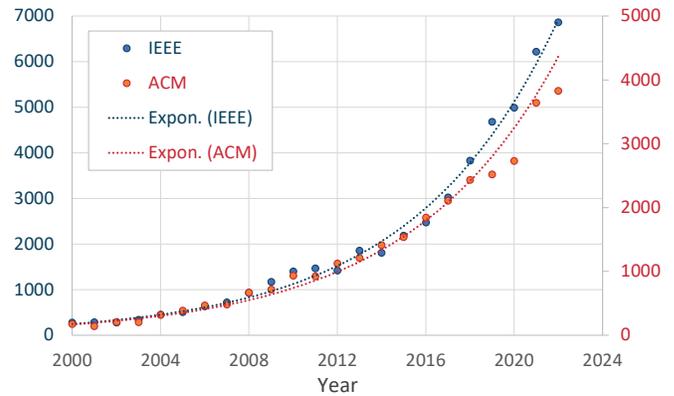


Fig. 10. An exponential growth in the number of IEEE and ACM publications related to emotion in images and videos in the recent two decades was observed. Statistics are based on querying the IEEE and ACM publication databases. The lower rate of expansion observed in the year 2020 can be attributed to the disruptions caused by the COVID-19 pandemic and the limitations it imposed on human subject studies.

image- and video-related publications. The growth in emotion-related research outpaced the baseline growth by a significant margin, suggesting that it has a higher future potential. The annual growth rate for the field of emotion in images and videos is 15-16%. If this growth continues, the annual number of publications in this field is expected to double every five years.

B. Modeling Evoked Emotion

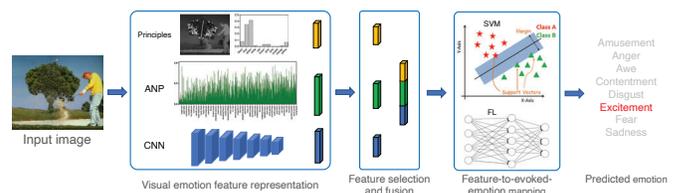


Fig. 11. A generalized framework for evoked emotion prediction from images.

Evoked emotions can be inferred when viewers see an image or a video, either based on changes in their physical body or the stimuli being viewed, i.e. explicit affective cues or implicit affective stimuli [114, 22, 69]. In this section, we will focus on implicit affective stimuli, particularly images and videos. Similar to other machine learning and computer vision tasks [182, 16], an evoked emotion recognition system from images or videos typically consists of three components: emotion feature extraction, feature selection and fusion, and feature-to-evoked-emotion mapping. The generalized frameworks for this process are illustrated in Figs. 11 and 12. The first step is to extract emotion features from the original images and videos, typically after pre-processing, and converting them to numerical representations that are easier to process. Feature selection and fusion aim to select discriminative features that are more relevant to emotions, reduce the dimensionality of features, and combine different types of features into a unified representation. Finally, each unified representation is classi-

fied. Feature-to-evoked-emotion mapping learns a classifier to project the feature representation to specific evoked emotions.

1) *Visual Emotion Feature Representation*: Prior to the emergence of deep learning, visual emotion features were primarily developed manually, drawing inspiration from various interrelated disciplines including computer vision, art theory, psychology, and aesthetics.

One straightforward approach employs existing *low-level vision features* as representations of emotion. Yanulevskaya et al. [183] extracted holistic Wiccest and Gabor features. Other researchers have employed scale-invariant feature transform (SIFT), histogram of oriented gradients (HOG), self-similarities, GIST, and local binary patterns (LBP) to extract low-level vision representations of emotion for video keyframes [72, 184]. Yang et al. [185] compared their method with multiple low-level vision features, including SIFT, HOG, and Gabor. Rao et al. [186] first employed bag-of-visual-words of SIFT features for each image block from multiple scales and then represented each block by extracting latent topics based on probabilistic latent semantic analysis. While such low-level vision representations are still used, such as LBP and optical flows [187], they do not effectively bridge the semantic gap and cannot bridge the more challenging affective gap [16].

More robust approaches to exploring visual emotion features employ *theories on art and aesthetics* to design and develop features. Art is often created with the intention of evoking emotional reactions from the audience. As Pablo Picasso claims, “Art washes away from the soul the dust of everyday life,” and as Paul Cézanne asserts, “A work of art which did not begin in emotion is not art,” [44, 188, 189]. Generally, art theory includes elements of art and principles of art. The elements of art, such as color, line, texture, shape, and form, serve as building blocks in creating artwork. Meanwhile, the principles of art correspond to the rules, tools, or guidelines for arranging and orchestrating the elements of art, such as balance, emphasis, harmony, variety, and gradation. Among these elements, color is the most commonly used artistic feature [44, 190, 191, 192, 193, 194, 187, 195], followed by texture. Lu et al. [46] investigated the relationship between shape and emotions through an in-depth statistical analysis. Zhao et al. [188] systematically formulated and implemented six artistic principles except for rhythm and proportion and combined them into a unified representation.

Aesthetics is widely acknowledged to have a strong correlation with emotion [182, 196]. Artwork that is aesthetically designed tends to attract viewers’ attention and create a sense of immersion. As early as 2006, Datta et al. developed computer algorithms to predict the aesthetic quality of images [52]. Among the various aesthetic features, composition, such as the rule of thirds, has been the most popular [52, 44, 196, 190, 191, 197]. The figure-ground relationship, which refers to cognitive feasibility in distinguishing the foreground and the background, and several other aesthetic features were designed [196]. Sartori et al. [198] to analyze the influence of different color combinations on evoking binary sentiment from abstract paintings. Artistic principles and aesthetics were used to organize artistic elements from different but correlated perspectives, which were sometimes not clearly differentiated

and extracted together [196, 197]. By considering relationships among different elements, artistic principles and aesthetics have been demonstrated to be more interpretable, robust, and accurate than artistic elements and low-level vision representations for recognizing evoked emotions [196, 188].

To bridge the gap between low-level visual features and high-level semantics, an *intermediate level of attributes and characteristics* was designed. These intermediate attributes and characteristics were then applied to the prediction of evoked emotions [199, 192, 193, 194]. For example, in addition to generic scene attributes, eigenface-based facial expressions were also considered [199]. These attributes performed better than low-level vision representations, but the interpretability was still limited.

High-level content and concept play an essential role in evoking emotions for images with obvious semantics, such as natural images. The number and size of faces and skins contained in an image have been used as an early and simple content representation [44]. Facial expressions of images containing faces are a direct cue for viewers to produce emotional reactions and are therefore often employed as content representation [192, 193]. Jiang et al. [72] developed a method that involved detecting objects and scenes from keyframes. Based on the observation that general nouns like “baby” were detectable but had a weak link to emotion, whereas adjectives like “crying” could provide strong emotional information but were difficult to detect, Borth et al. [200] introduced adjective noun pairs (ANPs) by adding an adjective before a noun, such as “crying baby.” The combination enabled strong ability to map concepts to emotions while remaining detectable. A large visual sentiment ontology named SentiBank was proposed to detect the probability of 1,200 ANPs. Thus, as a milestone, SentiBank cannot be ignored as a baseline in performance evaluation, even in the current deep learning era. A multi-lingual visual sentiment ontology (MVSO) [201] was later extended to deal with different cultures and languages. About 16K sentiment-biased visual concepts across 12 languages are constructed and detected. ANP representations are widely used as high-level semantic representations [72, 192, 202, 193, 194]. These content and concept representations achieve the best performance for the images containing such semantics but fail for abstract paintings.

Fortunately, the rise of deep and very deep Convolutional Neural Networks (CNNs) in image classification has led to deep learning becoming the primary learning strategy in various fields, including computer vision and natural language processing. This is also the case for evoked emotion modeling [16]. Given sufficient annotated data, deep learning models can be trained to achieve superior performance on various types of images, including natural and social photographs, as well as abstract and artistic paintings. In many cases, features are automatically learned without the need for manual crafting.

Recently, *global representation at the image level* has been demonstrated to hold promise for evoked emotion analysis. One approach to extract deep features is to directly apply pretrained CNNs, such as AlexNet [203], VGGNet [204], and ResNet [205], to the given images and obtain responses of the last (few) fully connected (FC) layers [206, 190, 71,

202, 207, 197, 194, 208, 209, 210, 211]. Other methods have begun to use the output of a transformer's encoder as visual reorientation [212, 213]. Xu et al. [206] demonstrated that the deep features from the last but one FC layer outperformed those from the last FC layer. The extracted CNN features were transformed to another kernelized space via discrete Fourier transform and sparse representation to suppress the impact of noise in videos [207]. Chen et al. [214] first obtained event, object, and scene scores with state-of-the-art pretrained detectors based on deep neural networks. They then integrated these into a context fusion network to generate a unified representation. More recently, Song et al. [195] used the pretrained MS Azure Cognitive Service API to extract objects contained in an image and corresponding confidence scores as visual hints, which were then transformed to TF-IDF representation. This pretrained deep feature can be essentially viewed as another handcrafted feature since the classifier used for the final emotion prediction needs to be separately trained. To enable end-to-end training, fine-tuning is widely employed to adjust the deep features to be more correlated with evoked emotions [215, 190, 71, 216]. Other inspiring improvements include multitask learning [217, 185] and emotion correlation mining [218, 219, 220] to better learn emotion-specific deep features. The image-level global representation extracts deep features from the global aspect taking the original image as input. To deal with the temporal correlations among successive frames in videos, 3-D CNN (C3D) [221] is adopted taking a series of frames as input [222]. Although contextual information is considered, the global representation treats local regions equally without considering their importance.

Strategies such as attention [223, 224, 225, 226] and sentiment maps [227, 228] have been widely used to extract *region-level representations* in order to emphasize the importance of local regions in evoked emotion prediction. Spatial attention is used to determine the correspondence between local image regions and detected textual visual attributes [223]. Besides operating on the global representation to obtain attended local representation, the spatial attention map is enforced to be consistent with prior knowledge contained in the detected saliency map [224]. Both spatial and channel-wise attentions are considered to reflect the importance of the spatial local regions and the interdependency between different channels [226]. In contrast, Fan et al. [225] investigated how image emotion could be used to predict human attention and found that emotional content could strongly and briefly attract human visual attention. Yang et al. [227] designed a weakly supervised coupled CNN to explore local information by detecting a sentiment-specific map using a cross-spatial pooling strategy, which only required image-level labels. The holistic and local representations were combined by coupling the sentiment map. To address the issue of overemphasis on local information and neglect of some discriminative sentiment regions, a discriminate enhancement map was recently constructed by spatial weighting and channel weighting [228]. Both the discriminate enhancement map and sentiment map were coupled with the global representation.

In recent years, a growing body of research has focused on developing effective *multilevel representations* [229, 230, 231,

219, 232, 233, 234, 235, 236, 237, 238]. One strategy has been to view different CNN layers as different levels [229, 231, 234]. Multiple levels of features were extracted at different branches (four [231] and four plus one main [229] branches). The features from different levels were then integrated by a Bidirectional Gated Recurrent Unit [229] or a fusion layer with basic operations like the mean. These two methods both claim that features from global to local levels can be extracted at different layers, but the correspondence between them is still unclear. This issue is partially addressed in the multilevel dependent attention network (MDAN) [234]. Based on the assumption that different semantic levels and affective levels are correlated, affective semantic mapping disentangles the affective gap by one-to-one mapping between semantic and affective levels. Besides the global-level learning at the highest affective level with emotion hierarchy preserved, local learning is incorporated at each semantic level to differentiate among emotions at the corresponding affective level [234]. Further, multihead cross channel attention and level-dependent class activation maps are designed to model level-wise channel dependencies and spatial attention within each semantic level, respectively. Further research is needed to explore more effective ways to map semantic and affective levels, particularly when the number of levels varies significantly.

Another strategy for extracting multilevel representations involves utilizing object, saliency, and affective region detection [219, 232, 236, 237, 238, 230, 235]. Inspired by the Stimuli-Organism-Response (S-O-R) emotion model in psychology, Yang et al. [219] selected specific emotional stimuli, including image-level color and region-level objects and faces, using off-the-shelf detection models. Corresponding to the selected stimuli, three specific networks were designed to extract the features, including CNN-based color and other global representations, Long-Short Term Memory (LSTM)-based semantic correlations between different objects, and CNN-based facial expressions. Besides the correlations between different objects based on graph convolutional network (GCN), Yang et al. [232] also mined the correlations between scenes and objects using a scene-based attention mechanism, motivated by the assumption that scenes guide objects to evoke distinct emotions. A similar approach was applied by Cheng et al., but with the input to the object detector being the extracted temporal-spatial features by C3D [238]. These methods treat objects, typically detected by Faster R-CNN, as regions, and reflect the importance of different regions through attention [219] or graph reasoning [232, 238]. Objects and faces were also respectively detected in [236] and [237], and corresponding CNN features were extracted. Rao et al. [230] employed an emotional region proposal method to select emotional regions and remove non-emotional ones. The region's emotion score was a combination of the probability of the region containing an object and that of the region evoking emotions. A similar approach for selecting emotionally charged regions was used by Zhang et al., but the emotion score was a weighted combination of two probabilities [235].

Efforts have been made to *combine handcrafted and deep representations* in order to take advantage of their complementary information. For example, Liu et al. [197] com-

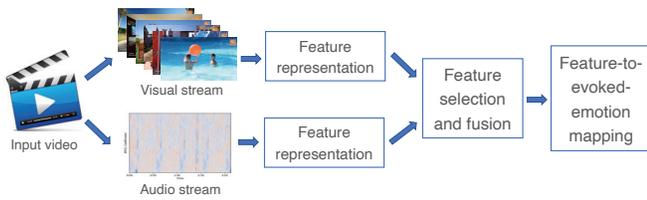


Fig. 12. A generalized framework for evoked emotion prediction from videos.

bined them after feature dimension reduction, whereas Chen et al. [239] used AlexNet-style deep CNNs [203] to train ANP detectors and achieved improved ANP classification accuracy. The DeepSentiBank has also been used to extract keyframe features in videos [240, 210]. More recently, the correlations of content features based on pretrained CNN and color features from color moments have been taken into consideration [241]. Specifically, the proposed cross-correlation model consists of two modules: an attention module to enhance color representation using content vectors and a convolution module to enrich content representation using pooled color embedding matrices. Further research is needed to investigate the potential of this approach to interact between handcrafted and deep representations.

2) *Audio Emotion Feature Representation*: The audio features used in video-based evoked emotion analysis are often sourced from the fields of acoustics, speech, and music signal processing [69]. Similar to visual features, audio features for emotion analysis can be divided into two categories: handcrafted and deep learning-based.

Among *handcrafted feature representations*, the Mel-frequency cepstral coefficients (MFCC) are commonly adopted features [72, 191, 184]. Energy entropy, signal energy, zero crossing rate, spectral rolloff, spectral centroid, and spectral flux are also employed [72, 184]. The openSMILE toolkit is a popular option for extracting handcrafted audio features [242]. El Ayadi et al. [243] classified speech features into four groups: continuous, qualitative, spectral, and Teager energy operator-based. Panda et al. [14] summarized music features into eight dimensions, including melody, harmony, rhythm, dynamics, tone color, expressivity, texture, and form. For more information on handcrafted audio feature representations, please refer to these papers.

Learning-based deep audio feature representations mirror the learning-based strategy used in the visual stream in videos, and some studies have explored the use of CNN-based deep features for audio streams. One approach is to send the raw audio signal directly to a 1-dimensional CNN. However, a more popular method involves transforming the raw audio signal into MFCC, which can be viewed as an image, before feeding it into a 2-D CNN [202, 237, 208, 238, 222, 211, 213]. When simultaneously considering the MFCC from multiple video segments, 3-D CNN or LSTM can be utilized. For further information on modeling the temporal information in videos, see Section IV-B4.

3) *Contextual Feature Representation*: The context information that is important in evoked emotion prediction can be divided into two categories: within visual media and across

modalities. For context within visual media, one common approach is to extract features from different levels that correspond to semantics at different scales, such as the global and local levels. For more details, see Section IV-B1. When considering context across modalities, the social context of users, including their common interest groups, contact lists, and similarity of comments to the same images, is taken into account in personalized image emotion recognition [193]. In addition to the visual and audio representations, motion representation is also considered [213].

4) *Feature Selection and Fusion*: As previously mentioned, features can be extracted within a single modality (e.g., image) or across multiple modalities. On the one hand, different types of features can have varying discriminative abilities. High-dimensional features may suppress the power of low-dimensional ones when fusing them together. High-dimensional features may cause a “curse of dimensionality” and corresponding overfitting. On the other hand, fusing different features together can improve the performance of emotion analysis by jointly exploiting their representation ability and complementary information. Therefore, feature selection and fusion techniques are often used, particularly for datasets with a small number of training samples.

Before being processed by the feature extractor, an image or video often undergoes preprocessing. Resizing an image to a fixed spatial resolution is straightforward, but videos pose a greater challenge due to differences in both spatial and temporal structures. It is important to determine how to combine frame-level or segment-level features into a unified video-level representation. Below, we first explore commonly used temporal information modeling in videos, followed by a summary of feature selection and fusion techniques.

In the area of *temporal information modeling*, there is a need to divide the input video into segments and extract keyframes to be used for feature extraction. A straightforward method is to use all frames of the entire video [207, 236], which ensures the least information loss but also results in high computational complexity. Common methods for segment sampling include fixed time interval [72, 184, 202, 237, 211], fixed number of segments [208, 222], and video shot detection [238, 210]. Sampling keyframes within a segment can be done through fixed frame sampling (uniform sampling) [72, 184, 237, 211], random sampling [208], middle frame selection [202], and mean histogram sampling (i.e., the frame with the closest histogram to the mean histogram) [191]. Recently, researchers have developed dedicated keyframe selection algorithms [240, 210]. One such method uses low-rank and sparse representation with Laplacian matrix regularization in an unsupervised manner, considering both the global and local structures [240]. Another method trains an image emotion recognition model on an additional image dataset to estimate the affective saliency of video frames [210]. Keyframes are selected based on the largest inter-frame difference of color histograms, after first sorting the segments (shots) based on affective saliency. Despite progress, effective and efficient selection of emotion-aware keyframes to enhance accuracy and speed remains an open challenge.

Another challenge is combining frame-level or segment-

level features into a unified video-level representation. Some straightforward methods include average pooling [72, 184, 207, 236], average pooling with temporal attention [208], max pooling [210], bag-of-words quantization [72, 184], LSTM [238, 222], and gated recurrent units (GRUs) [237]. Another approach is to perform emotion prediction at the segment level and then combine the results using methods such as majority voting [202, 211].

Feature selection, especially for handcrafted features, is often employed when extracted features have high dimensions that contain redundant, noisy, or irrelevant information, the dimensions of different kinds of features to be fused vary significantly, or the size of the training samples is small. Some commonly used feature selection methods in evoked emotion modeling include cross-validation on the training set [183], wrapper-based selection [44], forward selection [46], and principal component analysis (PCA)-based selection [44, 46, 188, 71]. Feature selection can be combined with feature fusion, particularly in the case of early fusion [188, 197]. If dimensions of the features being fused are similar, feature selection is carried out after fusion [188]. However, if dimensions are significantly different, feature selection is usually performed first [197] to prevent an overemphasis on high-dimensional features.

The integration of multiple features within a single modality or different kinds of representations across different modalities through *feature fusion* plays a crucial role in emotion recognition [22]. Model-free fusion, which operates independently of learning algorithms, is widely used and encompasses early fusion [44, 196, 188, 198, 190, 229, 225, 197, 230, 219, 232, 233, 195, 237, 208], late fusion [199, 194, 241, 202, 211], and hybrid fusion [210]. Hierarchical fusion that incorporates various feature sets at different levels of the hierarchy is also employed [187].

Model-based fusion, however, is performed explicitly during the learning process. Kernel-based fusion [72, 236] and graph-based fusion are often used for shallow learning models, whereas attention-based [235, 222, 213], neural network-based [184, 238], and tensor-based fusion strategies have been recently employed for deep learning models.

A Transformer encoder with multihead self-attention layers is used in sentiment region correlation analysis to exploit dependencies between regions [235]. The self-attention transforms original regional features into a higher-order representation between implied regions. Ou et al. used a local-global attention mechanism to explore the intrinsic relationship among different modalities and for different segments [222]. The local attention evaluates the importance of different modalities in each segment, whereas the global attention captures the weight distribution of different segments. In contrast, Pang et al. [184] employed the deep Boltzmann machine (DBM) to infer non-linear relationships among different features within and across modalities by learning a jointly shared embedding space. Cheng et al. [238] proposed an adaptive gated multimodal fusion model, which first mapped features from different modalities to the same dimension and then employed a gated multimodal unit (GMU) to find an intermediate representation. For a comprehensive survey on feature fusion strategies, please

refer to [22].

5) *Feature-to-Evoked-Emotion Mapping*: Based on the emotion representation models discussed in Section II-A, various evoked emotion analysis tasks can be undertaken, including classification, regression, retrieval, detection, and label distribution learning [22]. The former three can also be classified into visual media-centric dominant emotion analyses and viewer-centric personalized emotion analyses, while the latter two are typically visual media-centric. After obtaining a unified representation, shallow or deep learning-based methods can be employed to map the features to evoked emotions.

A variety of machine learning algorithms have been utilized to *learn the mapping* between the unified representation after feature fusion and evoked emotions. These algorithms include naïve Bayes [44, 200, 196], logistic regression [200, 199, 195, 184], SVM and support vector regression with linear and radial basis function (RBF) kernels [183, 46, 200, 199, 188, 197, 195, 72, 191, 207, 236, 240, 210, 187, 211], linear discriminant analysis (LDA) [211], multiple instance learning [186], random forest [195, 210], decision tree [195], ensemble learning [202], a mixture of experts [237], sparse learning [192, 194], and graph/hypergraph learning [193]. There is still room for innovation in designing emotion-sensitive mapping algorithms.

The most common *deep learning-based mapping* is a multi-layer perceptron based on one or more fully connected layers. The difference mainly lies in objective loss functions, such as cross-entropy loss for classification [215, 71, 217, 229, 216, 185, 227, 230, 231, 232, 241, 233, 228, 234, 212, 235, 240, 238, 210, 222], Euclidean and mean squared error loss for regression [190, 195], contrastive loss and triplet loss for retrieval, and Kullback-Leibler (KL) divergence loss for label distribution learning [217, 218, 189]. Unlike shallow learning methods, these deep learning techniques can typically be trained in an end-to-end manner. Some improvements have also been made to better explore the interactions between different tasks and the characteristics and relationships of different emotion categories, such as through the use of polarity-consistent cross-entropy loss and regression loss [226, 208] and hierarchical cross-entropy loss [219]. More details are given in the following.

The aforementioned evoked emotion analysis tasks are interconnected. For instance, the detection of emotional regions can inform the emotion classification task, and the emotion category with the largest probability in label distribution learning often aligns with the emotion classification results. However, a notable difference between evoked emotion modeling and traditional computer vision or machine learning lies in the existence of specific correlations among emotions. We summarize recent advancements in multitask learning and the exploration of emotion correlations in evoked emotion modeling.

Multi-task learning has been shown to significantly improve performance compared to single-task learning by leveraging appropriate shared information and imposing reasonable constraints across multiple related tasks [244]. It has become popular in evoked emotion modeling, especially when training data is limited for each task. Based on the relations

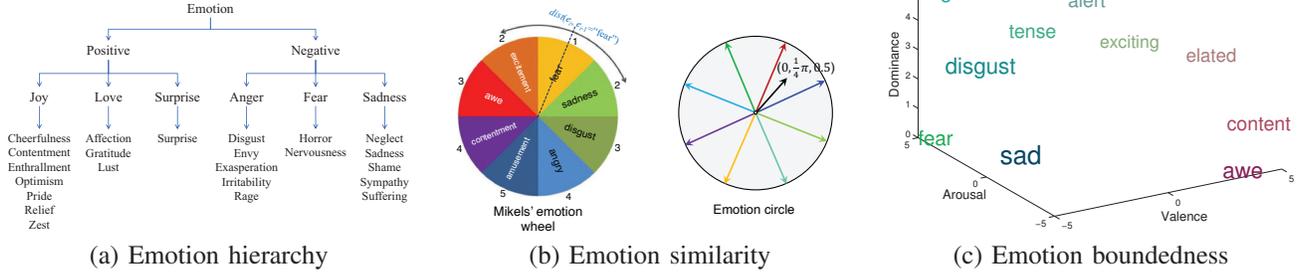


Fig. 13. Illustration of different emotion correlations. (a) Emotion hierarchy: a three-level hierarchy from sentiment to primary emotion and secondary emotion [234]. (b) Emotion similarity: Mikels’ emotion wheel on the left [193] and Emotion circle on the right [218]. (c) Emotion boundedness: rough transformation between discrete emotion categories and continuous emotion values in the VAD space [46].

explored, multitask learning can be categorized into task relation learning-oriented [217, 185, 227, 230, 233, 228], testing data relation learning-oriented [192], feature relation learning-oriented [194], and viewer relation learning-oriented approaches [193]. Different evoked emotion analysis tasks are often jointly performed, such as classification and distribution learning [217, 230], classification and retrieval [185], and classification and detection [227, 228, 233]. Joint optimization of different objective losses allows models to extract more discriminative feature representations. For example, multitask shared sparse regression was proposed to predict continuous emotion distributions of multiple testing images with sparsity constraints, which takes advantage of feature group structures [192]. Different constraints across features are considered [194] to reflect their importance by a weighting strategy, which can be viewed as a special late fusion category. Rolling multitask hypergraph learning is proposed to simultaneously predict personalized emotion perceptions of different viewers where social connections among viewers are considered [193].

Exploring correlations among different emotion categories or continuous emotion values can improve evoked emotion analysis. Commonly considered emotion correlations include:

- *Emotion hierarchy.* As emotion categories researchers attempt to model become more diverse and nuanced, level of granularity increases [234]. As shown in Fig. 13(a), emotions can be organized into a hierarchy, which has been exploited in evoked emotion analysis. By considering the polarity-emotion hierarchy, i.e. whether two emotion categories belong to the same polarity, polarity-consistent cross-entropy loss [208] and regression loss [226] are designed to increase the penalty of predictions that have the opposite polarity to the ground truth. Hierarchical cross-entropy loss has been proposed to jointly consider both emotion and polarity loss [219]. For each level in the emotion hierarchy, one specific semantic level is mapped with local learning to acquire corresponding discrimination [234].
- *Emotion similarity.* Similarities or distances between emotions can vary, with some being closer than others. For example, sadness is closer to disgust than it is to contentment. To account for these similarities, the Mikels’ emotion wheel was introduced [193] (Fig. 13(b) left). Pairwise emotion similarity is defined as the reciprocal

of “1 plus the number of steps required to discriminate one emotion from another”. Using the Mikels’ wheel, the emotion distribution is transformed from a single emotion category [217, 230]. Chain center loss is derived from the triplet loss, with anchor-related-negative triplets selected based on emotion similarity [220]. A more accurate method of measuring emotion similarity, based on a well-grounded circular-structured representation called the Emotion Circle, has also been designed [218] (Fig. 13(b) right). Each emotion category can be represented as an emotion vector with three attributes (i.e., polarity, type, and intensity) and two properties (i.e., similarity and additivity), allowing for vector addition operations.

- *Emotion boundedness.* Not every combination of valence, arousal, and dominance values make sense in an emotion space. The transformation between discrete emotion states and their rough continuous values is often possible [46, 16] (Fig. 13(c)). For example, positive valence is linked to happiness, whereas negative valence is linked to sadness or anger, even though exact boundaries may not be clear. When performing multitask learning involving both classification and regression, it is helpful to consider the constraints on these values. For example, it would not be valid to predict happiness with a negative valence. The BoLD dataset (Section III-A3) leverages this concept to check the validity of crowdsourced annotations [51].

6) *Preliminary Benchmark Analysis:* We provide a summary of image- and video-based evoked emotion classification accuracy of various representative methods, reported on the FI [71] and VideoEmotion-8 [72] datasets, respectively. The compared methods for image emotion prediction include SentiBank [200], artistic principles [188], DeepSentiBank [239], fine-tuned AlexNet [203], fine-tuned VGG-16 [204], fine-tuned ResNet-101 [205], progressive CNN [215], multilevel deep representation network (MldrNet) [231], weakly supervised coupled network (WSCNet) [227], polarity-consistent deep attention network (PDANet) [226], Multi-level region-based CNN (MlrCNN) [230], stimuli-aware visual emotion analysis network (SAVEAN) [219], scene-object interrelated visual emotion reasoning network (SOLVER) [232], multilevel dependent attention network (MDAN) [234], and SimEmotion [212]. Results in Table II indicate that: (a) fine-tuning deep neural networks, particularly those with more layers, outper-

TABLE II
PERFORMANCE COMPARISON OF REPRESENTATIVE METHODS FOR 8-CLASS EVOKED EMOTION CLASSIFICATION FROM IMAGES ON THE FI DATASET, MEASURED BY AVERAGE ACCURACY (%). *THE RESULTS ARE MAINLY OBTAINED FROM MDAN [234] AND SIMEMOTION [212].

Method	Venue	Accuracy (%)
SentiBank [200]	ACM MM 2013	44.5
Artistic principles [188]	ACM MM 2014	46.1
DeepSentiBank [239]	arXiv 2014	53.2
Fine-tuned AlexNet [203]	NeurIPS 2012	58.3
Fine-tuned VGG-16 [204]	ICLR 2015	65.5
Fine-tuned ResNet-101 [205]	CVPR 2016	66.2
Progressive CNN [215]	AAAI 2015	56.2
MldrNet [231]	NPL 2020	65.2
WSCNet [227]	CVPR 2018	70.1
PDANet [226]	ACM MM 2019	72.1
MlrCNN [230]	NEUCOM 2019	75.5
SAVEAN [219]	TIP 2021	72.4
SOLVER [232]	TIP 2021	72.3
MDAN [234]	CVPR 2022	76.4
SimEmotion [212]	TAFFC 2022	80.3

TABLE III
PERFORMANCE COMPARISON OF REPRESENTATIVE METHODS FOR 8-CLASS EVOKED EMOTION CLASSIFICATION FROM VIDEOS ON THE VIDEOEMOTION-8 DATASET, MEASURED BY AVERAGE ACCURACY (%). *THE RESULTS ARE MAINLY OBTAINED FROM VAANET [208] AND TAM [213].

Method	Venue	Accuracy (%)
SentiBank [200]	ACM MM 2013	35.5
E-MDBM [184]	TMM 2015	40.4
ITE [73]	TAFFC 2018	44.7
V.+Au.+At. [72]	AAAI 2014	46.1
CFN [214]	ACM MM 2016	50.4
V.+Au.+At.+E-MDBM [184]	TMM 2015	51.1
Kernelized [207]	TMM 2018	49.7
Kernelized+SentiBank [207]	TMM 2018	52.5
VAANet [208]	AAAI 2020	54.5
KeyFrame [210]	MTAP 2021	52.9
FAEIL [207]	TMM 2021	57.6
TAM [213]	ACM MM 2022	57.5

forms handcrafted features (e.g., fine-tuned ResNet-101 [205] versus SentiBank [200]); (b) exploring local feature representations and combining representations from different levels can increase accuracy to around 75% (e.g., MDAN [234]); and (c) SimEmotion, which uses large-scale language-supervised pre-training, achieves an overall accuracy of around 80%, which remains lower than traditional computer vision tasks. The results highlight the challenges in evoked emotion prediction due to large intra-class variance and the need for further progress toward human-level emotion understanding.

The compared methods for video emotion prediction include SentiBank [200], enhanced multimodal deep Boltzmann machine (E-MDBM) [184], image transfer encoding (ITE) [73], visual+audio+attribute (V.+Au.+At.) [72], CFN [214], V.+Au.+At.+E-MDBM [184], Kernelized features and Kernelized+SentiBank [207], visual-audio attention network (VAANet) [208], KeyFrame [210], frame-level adaptation and emotion intensity learning (FAEIL) [207], and temporal-aware multimodal (TAM) methods [213]. Results in Table III indicate that: (a) fusing information from multiple modalities is more effective than using a single modality (e.g.,

VAANet [208] versus SentiBank [200]); (b) the current best accuracy is less than 60%; and (c) effectively fusing information from different modalities and selecting key segments or frames are two challenges. Evoked emotion prediction from videos is even more challenging than image-based prediction and requires further research efforts.

C. Facial Expression and Microexpression Recognition

Facial expressions play a crucial role in natural human communication and emotion perception. Facial expression recognition (FER) involves automatic identification of a person’s emotional state through analysis of images or video clips and has been a long-standing research topic in computer vision and affective computing.

1) *Earlier Approaches:* Before 2012, traditional handcrafted features and pipelines were commonly used. The process generally included the following steps: detecting facial regions, extracting handcrafted features (e.g., LBP [245], non-negative matrix factorization (NMF) [246], HOG [247]) from facial regions, and employ a statistical classifier (e.g., SVM [248]) to recognize emotions.

Readers interested in traditional methodologies are advised to refer to survey articles [249, 250, 251, 252].

Since the creation of the ImageNet dataset and deep CNN AlexNet in 2012, DNNs have demonstrated remarkable image representation capabilities. FER researchers have established large-scale datasets (e.g., EmotoNet [253], AffectNet [78]) that provide ample training data for DNNs. Consequently, deep learning approaches have become the dominant approach in FER. A survey by Li and Deng [4] provides a comprehensive summary of deep learning-based FER methods from 2012 to 2019. During this period, researchers proposed several DNN techniques to improve FER performance, which the authors categorize as follows:

- Adding auxiliary blocks to the typical backbone network (e.g., ResNet [205] and VGG [204]). The Scoring ensemble (SSE) [254] proposed three auxiliary blocks to extract features from the shallow-, intermediate-, and deep-layers, respectively.
- Ensembling different models to achieve outstanding performance. For example, [255] concatenated features from three different networks—VGG13 [204] and VGG16 [204], and ResNet [205].
- Designing specialized loss functions (such as Center loss [256], Island loss [257], and (N+M)-tuple cluster loss [258]) to learn facial features.
- Leveraging multitask learning to learn various features from facial images. For instance, [259] proposed MSCNN to jointly learn face verification and facial expression recognition tasks.

Moreover, some studies concentrated on FER from video clips using spatiotemporal networks to capture temporal correlations among video frames. Some researchers [260] also used recurrent neural networks (RNNs) (e.g., LSTM), whereas others [261, 262] used 3-D convolutional networks such as C3D [221].

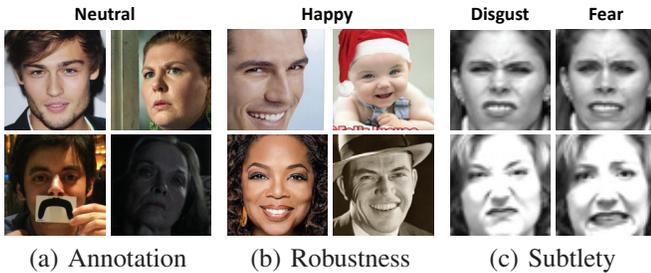


Fig. 14. Three challenges in facial expression recognition. Images in (a-b) are obtained from the AffectNet dataset [78] and images in (c) are obtained from the CK+ dataset [76].

2) *Recent Approaches*: Although deep learning-based methods have achieved remarkable success, FER remains a difficult task due to these three challenges.

- *Annotation*. Annotations of FER datasets contain much ambiguity. Each annotator subjectively evaluates facial expressions in images, leading to different annotations to the same image. Some images also have inherent ambiguity, making it difficult to assign a clear emotion label. Low image quality (such as occlusion or poor lighting) and ambiguous facial expressions exacerbate this problem. For instance, the images in Fig. 14 (a) are labeled as neutral, but this label is uncertain due to annotator subjectivity and/or image quality.
- *Robustness*. FER contains several sources of disturbance. Datasets are made up of individuals of varying ages, genders, and cultures. In addition, visual variations (such as human facial pose, illumination, and occlusions) commonly exist in facial images, which cause distinct appearance differences. These identity and visual variations can make it difficult for FER models to extract useful features. As illustrated in Fig. 14 (b), image appearances can vary greatly due to differences in age, gender, race, facial pose, and lighting. Moreover, individuals with different identities and cultures might display their facial expressions differently, adding another challenge for FER models.
- *Subtlety*. Some facial expressions are delicate, and some emotions can also be conveyed through subtle facial actions. Such fine distinctions can make it challenging to distinguish between emotions. For instance, the difference between “fear” and “disgust” images in Fig. 14 (c) is nuanced. Thus, FER models must efficiently extract discriminative features to differentiate emotions.

Most deep learning methods developed after 2019 have attempted to address these problems. Regarding the annotation problem, Zeng et al. pointed out the inconsistency of emotion annotation across different FER datasets due to annotators’ subjective evaluations of emotion [263]. To address this issue, they proposed a framework called Inconsistent Pseudo Annotations to Latent Truth (IPA2LT), which trains multiple independent models on different datasets separately. These models may assign inconsistent pseudo labels to the same image because each model reflects the subjectivity of the corresponding dataset annotators. By comparing image’s inconsistent labels,

IPA2LT estimates the latent true label. Another factor that can contribute to annotation ambiguity is uncertain facial images, such as blurry images or those with ambiguous emotions. To mitigate the effect of uncertain images, researchers have proposed various approaches. Wang et al. developed a Self-Cure Network (SCN) to identify and then suppress uncertain images [264]. SCN uses a self-attention mechanism to estimate the uncertainty of facial images and a relabeling mechanism to adjust labels of those images. Chen et al. argued that annotating uncertain images with multilabel and intensity, rather than the one-hot label commonly used in current FER datasets, is more suitable [265]. Label Distribution Learning (LDL) allows models to learn label distribution (i.e., multilabel with intensity). Chen et al. introduced an approach called Label Distribution Learning on Auxiliary Label Space Graphs (LDL-ALSG) [265]. Given one image, LDL-ALSG first leverages models of related tasks (such as AU detection) to find its neighbor images. Then LDL-ALSG employs a task guide loss to let images learn a similar label distribution (i.e., multilabel with intensity) with the neighbors. She et al. combined LDL and uncertain image estimation. The proposed DMUE network mines the label distribution and estimates the uncertainty of images together [266].

To address robustness challenges in FER, researchers have explored mitigating the impact of identity variations on recognition [267, 268]. Chen and Joo investigated the influence of gender and revealed that models tend to recognize women’s faces as happy more often than men’s, even when smile intensities are the same [267]. To overcome this issue, the authors proposed a method that first detected facial AUs and then applied a triplet loss to ensure that people with similar AUs exhibited similar expressions, regardless of gender [267]. Zeng et al. demonstrated that emotion categories can introduce bias into a dataset [268]. In some datasets, emotions (such as disgust) occur less frequently than more prevalent emotions such as happiness and sadness, leading to poor performance of networks trained on these datasets on minority emotion classes. The authors utilized a million-image-level facial recognition dataset (much larger than the FER dataset) and a meta-learning framework to address the issue [268]. Li et al. recognized the category imbalance challenge and addressed it by proposing AdaReg loss to dynamically adjust the importance of each category [269].

Other researchers have focused on visual disturbance variations, which lead to the robustness issue. Wang et al. designed a region attention network to capture important facial regions, thus obtaining occlusion-robust and pose-invariant image features [270]. Zhang et al. used a Deviation Learning Network (DLN) to learn identity-invariant features [271]. Wang et al. considered both identity and pose variations together [272], where an encoder was followed by two discriminators that classified pose and identity independently, while the encoder extracted features that were invariant to both. A classifier for expressions was then used to produce expression predictions.

To address the subtlety problem in FER, Ruan et al. decomposed facial expression features into shared features that represented expression similarities, and unique features that represented expression-specific variations using a Feature

Decomposition Network (FDN) and a Feature Reconstruction Network (FRN), respectively [273]. The authors further addressed robustness and subtlety problems by proposing a disturbance feature extraction model (DFEM) to identify disturbance features (such as pose and identify) and an adaptive disturbance-disentangled model (ADDM) to remove disturbance features extracted by DFEM and extract the discriminative features across different facial expressions. Farzaneh and Qi enhanced Center Loss [256, 274]. Although Center Loss can learn discriminative features, it can also include some irrelevant features. The proposed Deep Attentive Center Loss adopts an attention mechanism to adaptively select important discriminative features. Xue et al. leveraged Transformers to detect discriminative features and showed that the original vision transformer (ViT) [276] can only capture most discriminative features but neglects other features [275]. Thus this work proposes the Multi-Attention Dropping (MAD) technique to randomly drop some attention maps, such that the network can characterize all comprehensive features except the most discriminative ones. Furthermore, Savchenko et al. [277] achieved remarkable performance on FER using the well-performing network EfficientNet [278] to extract discriminative features.

TABLE IV
PERFORMANCE COMPARISON OF REPRESENTATIVE METHODS FOR 7-CLASS FACIAL EXPRESSION RECOGNITION ON THE AFFECTNET DATASET, MEASURED BY AVERAGE ACCURACY (%).

Method	Venue	Accuracy (%)
IPA2LT [263]	ECCV 2018	57.31
IPFR [272]	ACM MM 2019	57.40
RAN [270]	TIP 2020	52.97
SCN [264]	CVPR 2020	60.23
DAFL [274]	WACV 2021	65.20
DMUE [266]	CVPR 2021	63.11
KTN [269]	TIP 2021	63.97
TranFER [275]	CVPR 2021	66.23
Face2Exp [268]	CVPR 2022	64.23
ADDL [279]	IJCV 2022	66.20
EfficientNet-B2 [277]	TAFRC 2022	66.34

3) *Preliminary Benchmark Analysis*: Table IV provides a concise overview of the recent FER method performance on the AffectNet dataset, one of the most comprehensive FER benchmarks. The table shows the emotion classification accuracy of selected representative methods, which were previously discussed in the text. As seen in the table, TranFER, ADDL, and EfficientNet-B2 are among the top-performing methods on the AffectNet benchmark. The utilization of advanced network structures from general image recognition tasks in TranFER and EfficientNet-B2 highlights the importance of drawing knowledge and expertise from the image recognition field in FER. However, it is important to note that the highest reported accuracy remains below 70%. Given the over 90% top-1 accuracy achieved by state-of-the-art image recognition methods on the challenging ImageNet benchmark, this deficit suggests significant potential for improvement in FER.

4) *Microexpression Recognition (MER)*: The methods described above are for recognizing facial expressions. However, individuals may consciously exhibit certain facial expressions

to conceal their authentic emotions. In contrast to conventional facial expressions, which can be deliberately controlled, microexpressions are fleeting and spontaneous and can uncover an individual's genuine emotions. A microexpression is brief in duration and can be imperceptible with the naked eye. MER often requires high frame-rate videos as input and the development of spotting algorithms to isolate temporally the microexpression within videos. A recent survey gives a more in-depth introduction to developments in MER [7].

D. Bodily Expressed Emotion Understanding (BEEU)

In everyday life, people express their emotions through various means, including via their facial expressions and body movements. Recognizing emotions from body movements has some distinct advantages over recognizing emotions from facial images for many computer and robotic applications:

- In crowded environments where facial images may be obscured or lack sufficient resolution, body movements and postures can still be reasonably estimated. This context is particularly important in robotic applications where the robot may not be close to all individuals in its environment.
- Due to privacy concerns, facial information may be inaccessible. For example, in some medical applications, sharing of facial images or videos is restricted to protect sensitive patient identity information.
- Incorporating body expressions as an additional modality can result in more accurate emotion recognition compared to using facial images alone. For example, when the person is not facing the camera, the camera cannot obtain a frontal view.

Psychologists have conducted extensive studies to examine the relationship between body movements and emotions. Research suggests that body movements and postures are crucial for understanding emotion, encoding rich information about an individual's status, including awareness, intention, and emotional state [280, 281, 282, 283, 114]. Several studies, including one published in *Science*, found that the human body may be more diagnostic than the face for emotion recognition [283, 284, 285].

However, the field of BEEU in visual media has progressed relatively slowly. Unlike FER, which has seen significant progress with deep learning methods since 2013, most BEEU studies relied on traditional, handcrafted features until 2018. The bottleneck for BEEU is the scarcity of large-scale, high-quality datasets. As mentioned in Section III-A3, collecting and annotating a dataset of bodily expressions with high-quality labels is extremely challenging and costly. Understanding and perception of emotions from concrete observations is heavily influenced by context, interpretation, ethnicity, and culture. There is often no gold-standard label for emotions, especially bodily expressions. Prior to 2018, research on bodily expression was limited to small, acted, and constrained lab-setting video data [286, 287, 288, 289]. These datasets were insufficient for deep learning-based models that required a large amount of data. The recent BoLD dataset by Luo et al. [51] introduced in Section III-A3 is the largest BEEU

dataset to date. It contains over 10,000 video clips of body movements with high-quality emotion labels. In addition, Randhavane et al. [74] proposed the E-Walk data, including 1,136 3-D pose sequences extracted from raw video clips. As a result, computer vision research is increasingly focused on emotion recognition through body movement recordings.

Because static body images alone can hardly inform a person’s emotions, BEEU often considers human video clips. According to input modality of the models, BEEU methods can be classified as pixel-based and skeleton-based. Pixel-based methods use entire video clips, whereas skeleton-based methods first extract 2D/3D pose information and then feed it into the models. Some BEEU works, which focus on movement when walking, are known as gait-based. Herein, we refer to them as skeleton-based as well because gait is also a 2D/3D pose. Figs. 15 and 16 illustrate the two kinds of methods.

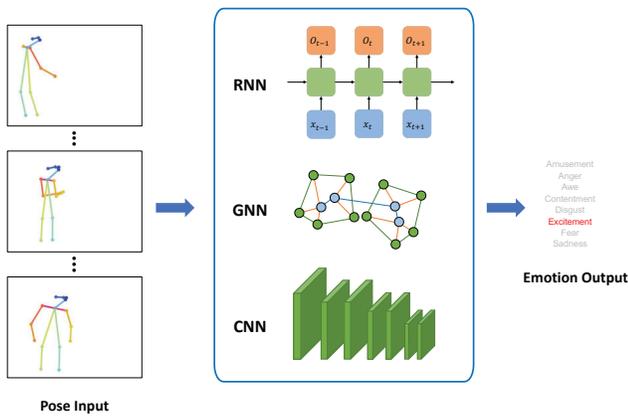


Fig. 15. Skeleton-based BEEU takes the 2D/3D pose sequence of the target human instance as input.

1) *Skeleton-based Methods*: Our review of past publications shows a greater adoption of skeleton-based approaches compared to pixel-based. This is due to two reasons. First, skeleton data, consisting of sequences of 2D/3D joint coordinates, requires less engineering effort to process than video clips. A straightforward method is to feed coordinates into a machine learning classifier (e.g. SVM) for direct emotion prediction results. Second, improved MoCap systems enable researchers to easily collect accurate 3-D poses from individuals walking in laboratory settings.

In early stages of skeleton-based BEEU research, a conventional approach was followed, which entailed the extraction of low-level features from 2D/3D pose sequences and subsequent utilization of a machine learning classifier to predict emotion. These features were categorized as follows:

- Frequent domain features, which transform temporal information into a frequency domain, are obtained through Fourier transformation. For instance, Li et al. used Fourier transformation to convert 3-D pose sequences into frequency domain features, which were then classified using linear discriminant analysis, naïve Bayes, decision tree, and SVM algorithms [290].
- Motion features, characterize the movement of body joints, include joint velocity and acceleration [51].

- Geometry features, which describe self-transformation of the body, encompass angles of specific skeletons and distance between certain joints, among others. In particular, Crenn et al. combined motion features, geometry features, and Frequent domain features into an SVM classifier [291].

Luo et al. compared traditional and deep learning-based methods using the BoLD dataset and developed the Automated Recognition of Bodily Expression of Emotion (ARBEE) system [51]. A traditional machine learning approach was designed, where motion and geometry features were extracted from 2-D pose sequences, and a random forest classifier was employed to categorize emotions. In addition, a ST-GCN was utilized to train and evaluate the BoLD dataset. The result indicated that carefully designed traditional machine learning methods outperformed the ST-GCN model from scratch.

Since the development of ARBEE, various deep learning-based methods have emerged. These methods can be broadly categorized into three groups based on the type of neural network used: RNN, Graph Neural Network (GNN), and CNN.

Randhavane et al. used a LSTM network, which is a type of RNN, to extract temporal features from a 3-D pose sequence, which were then concatenated with handcrafted features for classification [74]. Bhattacharya et al. leveraged a semi-supervised technique to improve the performance of an RNN model [292]. The work consisted of a GRU (a kind of RNN model) for feature extraction from a 3-D pose sequence, followed by an autoencoder with both encoder and decoder components. During training, when the input data was labeled with emotions, the classifier after the encoder produced the emotion prediction, and the decoder reconstructed the 3-D pose. If the input lacked emotion labels, only the decoder was used for reconstruction.

Bhattacharya et al. [293] adopted ST-GCN [294] to classify emotion categories from 3-D pose sequences. To increase the size of the training set, the authors used a Conditional Variational Autoencoder (CVAE) to generate some synthetic data. Banerjee et al. combined GCN and natural language processing (NLP) techniques to achieve zero-shot emotion recognition, which entailed recognizing novel emotion categories not seen during training [295]. The authors used ST-GCN to extract visual features from the 3-D pose sequences and used the word2vec method to obtain word embeddings from emotion labels. An adversarial autoencoder was used to align visual features with word embeddings. During inference, the system searched for the emotion label that best matched the output visual feature.

Inspired by the success of image recognition, some studies have attempted to convert skeleton sequences into images. Narayanan et al. embedded 3-D pose sequences into images, then utilized a CNN for classification [296]. Hu et al. employed a two-stream CNN, where one stream directly embedded the 3-D pose into an image, and the other stream converted handcrafted features from the 3-D pose into another image [297]. The two CNNs were integrated using Transformer layers.

2) *Pixel-based Methods*: Pixel-based networks for BEEU require video clips as input. Due to an increased redundancy

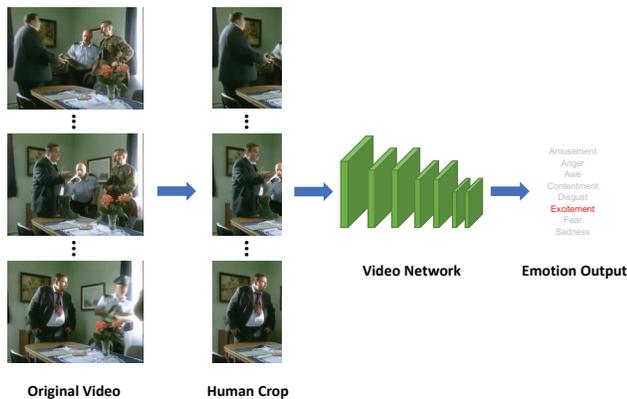


Fig. 16. Pixel-based BEEU uses raw videos as input. The target human is cropped and the resulting frames are input into a video analysis neural network.

in video clips compared to 2D/3D pose sequences, these networks necessitate larger training datasets to extract distinctive features. With the availability of the large-scale BEEU dataset, researchers have focused on pixel-based methods. Because the well-established field of action recognition, which also uses videos to analyze human behavior, has many similarities to BEEU, current BEEU research often incorporates networks from action recognition.

The ARBEE study benchmarked the performance of various action recognition networks on the BoLD dataset [51]. They cropped human body regions from raw videos as the input and then fed them into different methods, including a traditional handcrafted feature method and three deep learning methods (Two-Stream (TS) [298], I3D [138], and TSN [299]). Both TS and TSN consisted of a two-stream network, where one stream processed the RGB image sequence and the other processed the optical flow sequence. TS and TSN used 2-D convolutional networks, whereas I3D employed a 3-D convolutional network. Results indicated that all deep learning methods significantly outperformed the traditional handcrafted feature method. Among the three deep learning methods, I3D performed worse than TS and TSN, potentially due to its requirement for more training data to reach a comparable level of performance. The BoLD dataset is not as extensive as action recognition datasets like Kinetics-400 [300], and thus, I3D trained on BoLD could not fully exhibit its capability as it could on Kinetics-400. At the same time, TS and TSN produced similar results.

The ARBEE study also evaluated the impact of a person’s face on BEEU model performance [51]. They designed an ablation study with three different video inputs: the whole body (crop the whole human body from the raw video clips), just the face (only crop the face part), and the body without the face (crop the whole human body but mask the facial region) [51]. Results showed that using either the face or body alone was comparable to using the whole body. This demonstrated that both the face and body contributed significantly to the final prediction. Although the whole body setting of the TSN model outperformed the separate models, it did so by combining facial and body emotions.

Most BEEU research has followed ARBEE to continue pixel-based approaches. Because ARBEE cropped human body regions as input, a direct idea to improve upon ARBEE is to explicitly utilize facial images and background images as input as well. Recent studies [301, 302] used an extra network to extract context information from whole images, then fused the context features with features extracted from body images. Some research adopts an additional network with facial images as input [303]. Moreover, inspired by the cutting-edge vision-language research of CLIP [304], Zhang et al. developed EmotionCLIP, a contrastive vision-language pretraining paradigm to extract comprehensive visual emotion representations from whole images, encompassing both context information and human body information simultaneously [305]. Because EmotionCLIP uses only uncensored data, it addresses the challenge of data scarcity in emotion understanding. EmotionCLIP outperforms state-of-the-art supervised visual emotion recognition methods and competes with many multimodal approaches across various benchmarks, demonstrating its effectiveness and transferability.

Certain studies delve deeply into the analysis of body gestures and movement in the context of BEEU. ARBEE currently provides only a generalized emotion label for entire video clips, lacking specific descriptions for human body gestures or movements. In contrast, Liu et al. and Chen et al. introduced datasets for *micro-gesture* understanding and emotion analysis, featuring detailed body gesture labels, such as crossed fingers, for each human movement clip [75, 306]. These enriched datasets have the potential to improve machines’ understanding of emotions conveyed through gestures. As mentioned in Section III-E, LMA is a comprehensive method for describing human movement. Wu et al. further advanced BEEU by presenting an LMA dataset that provides accurate LMA labels for human movements [307]. Wu et al. incorporated a novel dual-task model structure that simultaneously predicts emotions and LMA labels, achieving remarkable performance on the BoLD dataset.

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT METHODS FOR BODILY EXPRESSIVE EMOTION UNDERSTANDING ON THE BoLD DATASET.

Method	Venue	mAP (%)
<i>Skeleton-based:</i>		
ST-GCN [263]	AAAI 2018	12.63
Random Forest [51]	IJCV 2020	13.59
<i>Pixel-based:</i>		
TS [298]	NeurIPS 2014	17.04
TSN [299]	ECCV 2016	17.02
I3D [138]	CVPR 2017	15.34
Filintisis et al. [302]	ECCVW 2020	17.96
Pikoulis et al. [303]	FG 2021	21.87
EmotionCLIP [305]	CVPR 2023	22.51
Wu et al. [307]	arXiv 2023	23.09

3) *Preliminary Benchmark Analysis and Current Directions:* Table V presents the results of various BEEU methods using the BoLD dataset, with performance measured by mean average precision (mAP) across 26 emotional categories. Results indicated that pixel-based methods outperformed skeleton-based ones, which was not surprising given

RGB images contained more information. Despite significant progress in BEEU, performance remains relatively low, with mAP scores below 25%.

As BEEU is a relatively new area in computer vision, we explore its potential future directions inspired by the trajectory of related areas like action recognition and FER. First, existing bodily expression datasets are not sufficiently large. BoLD contains only thousands of instances, which is much less than contained in action recognition datasets such as Kinetics-400. The BoLD team is developing larger and more comprehensive datasets to satisfy data requirements of deep learning methods, but those expansions are not yet complete. Second, existing approaches are largely based on action recognition methods without leveraging deep affective features. Previous work only applied low-level features, neglecting characteristics of concurrent body movement. Third, annotation ambiguity, similar to FER, is also challenging. Fourth, any method that segments body regions must account for changing relationships among body sections (i.e., Shape Change in LMA) which is crucial for bodily expressed emotion, and may be particularly relevant to the dimensional emotion model because Shape Change reveals approach and retreat. Section V discusses technological barriers more broadly and in greater depth.

E. Integrating Multiple Visual Input to Model Expression

We discussed earlier how facial and body images have been used as separate inputs to identify human emotions. As was already established, the context of people in a scene also contributes to inferring their emotion. Humans synthesize all visual information to produce emotion determinations. Naturally, computers are also capable of making an identification by combining all visual inputs including facial images, body images, and context information.

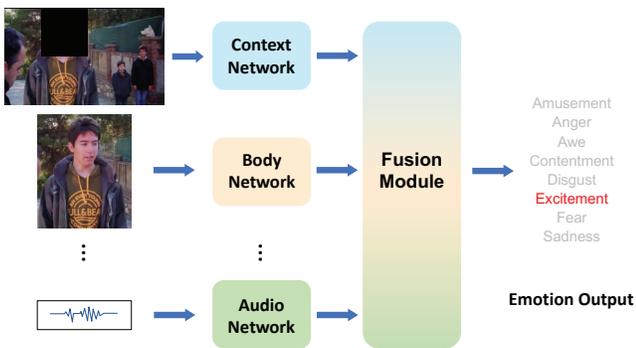


Fig. 17. The multistream network extracts features from different inputs, which can be various visual inputs (such as entire and body images) or inputs from other modalities (such as audio). The fusion module combines those features.

As illustrated in Fig. 17, recent work used multistream networks to fuse different visual inputs. Kosti et al. [136] and Lee et al. [81] developed initial approaches by adopting a two-stream network. Specifically, Kosti et al. used body images and entire images as input to extract context information and body features separately. Lee et al. adopted faceless images (i.e., cropped the human face out of the entire image) and facial

images as input, and a feature fusing network that dynamically fused context and facial features based on their significance.

Subsequent research has taken two paths. Some researchers have explored more kinds of visual inputs. For example, Mittal et al. demonstrated that a depth map can indicate how people interact with each other [137]. The estimated depth map from the raw image served as one input. In addition, this work used three other inputs—the facial image, the 2-D pose, and the bodiless (i.e., cropping the human body out of the entire image) images—to extract facial information, body posture information, and context information separately. This was a four-stream network overall. Studies [301, 302, 303] have also used a three-stream network with the entire, facial, and body images as input.

Other researchers focused on more effective fusion of visual features. Existing multistream approaches in typical emotion recognition adopt a simple fusion strategy, i.e., ensemble the predictions from each stream [301, 302, 303]. To improve upon this, Le et al. proposed the Global-Local Attention (GLA) module to enhance interaction between facial and context features [308].

F. Multi-modal Modeling of Expressions

Emotions can be conveyed and perceived not just through visual signals, but also through text and audio. To effectively process multimodal signals, however, is nontrivial. With advancements in deep learning, new technical approaches to this problem have emerged, particularly the vision-and-language model [309, 304]. Similar to methods discussed in Section IV-E, multimodal approaches also follow a multi-stream pipeline, as illustrated in Fig. 17. These techniques use independent networks to extract features from different inputs of multiple modalities and fuse these features with a fusion network.

Multi-modal approaches typically utilize backbone networks, such as BERT [310] and ResNet, to extract text and audio features. The visual feature extraction process, however, differs and is typically performed using one of the following three methods:

- *Region features.* A detection network is employed to identify regions of interest (ROIs), from which features are extracted.
- *Grid features.* A backbone network, such as ResNet 101, is used to extract features for the entire image.
- *Patch projection.* The image is split into patches, and a linear layer is used to generate a liner embedding, as described in ViT [276].

For the fusion process, a simple approach is to directly ensemble the final prediction of the different networks. Another method involves concatenating feature maps from different modality networks and using a single network to fuse the features. Yet another approach is to use multiple networks to process different features, with interactions between the different networks.

Some research in emotion recognition focuses on integrating audio and visual inputs. For example, Tzirakis et al. attempted to fuse audio signals with facial images [311].

They incorporated an audio stream network to extract audio features, followed by an LSTM network to fuse facial and audio features. Antoniadis et al. combined audio signals with multiple visual modalities, including facial image, body image, and context image [312]. In contrast to using CNN, Shirian et al. utilized a GCN structure to process audio, facial, and body inputs, and then used a pooling layer to process the fused features [313].

With the rapid advancement of NLP, there has been a surge in research on multimodal sentiment analysis, which incorporates text, audio, and visual input. Survey papers by Gandhi et al. [70] and Zhao et al. [22] provided comprehensive overviews of advancements in multimodal sentiment analysis techniques. Noteworthy methods from recent years include the development of a multitask network by Akhtar et al. that used text, acoustic, and visual frames of videos as input with inter-modal attention modules to adjust the contribution of each modality [314]. Most existing methods adopt multimodal emotion labels as the supervision, ignoring unimodal labels. Yu et al. proposed a label generation module that generated unimodal labels for each modality in a self-supervised manner [315], enabling a multitask network to train all unimodal and multimodal tasks simultaneously. Jiang et al. adopted several baseline models for each modality input and used PCA to find the optimal feature for the modality [316]. Further, an early fusion strategy combined all features. Yang et al. and considered the consistency and differences among various modalities in a unified manner, using common and private encoders to learn modality-invariant and modality-unique features, respectively, across all modalities [317]. With a similar motivation, Zhang et al. proposed a cascade and specific scoring model to represent the inter- and intra-relationship across different modalities [318]. Zhang et al. used reinforcement learning and domain knowledge to process fused features from multiple modalities in conversational videos [319]. Through reinforcement learning, Dueling DQN predicted the sentiment of the target sentence based on features of previous sentences. Moreover, information in the first several sentences was used as domain knowledge for subsequent predictions. Mittal et al. incorporated multiple visual inputs, resulting in the inclusion of five distinct types of input: face, body, context, audio, and text [320].

To show recent advancements in multimodal sentiment analysis, Table VI presents the performance of selected representative methods on the widely used benchmark, the MOSEI dataset. Performance is evaluated through mean absolute error (MAE) and binary accuracy (Acc-2) metrics. The top-performing methods were Self-MM [315] and FDMER [317]. Self-MM utilizes a multitask network to train all modalities simultaneously, whereas FDMER employs modules to capture shared and individual features across modalities. The essence of multimodal sentiment analysis remains the optimization of relationships between different modalities.

In addition, some research centers on fusing multimodal features. For example, Cambria et al. presented Sentic Blending to fuse scalable multimodal input [321]. In a multidimensional space, it constructed a continuous stream for each modality, which depicted the semantic and cognitive development of

TABLE VI
PERFORMANCE COMPARISON OF REPRESENTATIVE METHODS FOR MULTIMODAL SENTIMENT ANALYSIS ON THE MOSEI DATASET, *THE RESULTS ARE MAINLY OBTAINED FROM FDMER [317].

Method	Venue	MAE	Acc-2 (%)
TFN [322]	EMNLP 2017	0.593	82.5
LMF [323]	ACL 2018	0.623	82.0
MFM [324]	ICLR 2019	0.568	84.4
ICCN [325]	AAAI 2020	0.565	84.2
MISA [326]	ACM MM 2020	0.555	85.5
Self-MM [315]	AAAI 2021	0.530	85.2
FDMER [317]	ACM MM 2022	0.536	86.1

humans. The streams of different modalities were then fused over time.

V. MODELING EMOTION: SIGNIFICANT TECHNOLOGICAL BARRIERS

Despite advancements propelled by deep learning and big data, solving the problem of emotion modeling is nowhere in sight. In this section, we share insights on some of the most significant technological barriers in computer vision (Section V-A), statistical modeling and machine learning (Section V-B), AI (Section V-C), and emotion modeling (Sections V-D, V-E, V-F, and V-G) that hinder progress in the field. At present, there are no straightforward solutions to these issues. Some fundamental technologies in multiple related fields must be further developed before substantial progress can be made in emotion understanding. Emotional intelligence plays a significant role in our cognitive functions such as decision-making, information retrievability, and attention allocation. Artificial emotional intelligence (AEI) will likely become an integral part of future-generation AI and robotics. To put it succinctly, AEI based on visual information is a “Holy Grail” research problem in computing.

A. Fundamental Computer Vision Methods

1) *Pre-training Techniques*: The task of annotating emotion is a time-consuming process. Acquiring a large-scale dataset for FER or BEEU on the level of ImageNet is difficult. To completely overcome the challenge of emotion recognition, DNNs must acquire sufficient representation capability from extensive datasets. This gap can be bridged by using pretraining techniques, in which a model is initially trained on a massive dataset for an upstream task and then fine-tuned for a downstream task. During pretraining, the model is expected to learn emotion recognition capabilities from the upstream task. The choice of upstream task is critical because it determines the amount of emotion-related capabilities the model can acquire.

The widely adopted pretraining strategy is to train the model on ImageNet for image classification tasks. However, the ARBEE team has demonstrated that this approach does not significantly enhance the performance of the BEEU task [51].

Our hypothesis is that image classification and emotion recognition require different types of discriminative features to be extracted. This observation reflects the fact that whereas most individuals can recognize common objects, a portion of

the population is unable to discern subtle emotions. In light of the current state of computer vision, two upstream tasks have been promising for pretraining: self-supervised learning (SSL) [327, 328] and image-text retrieval [304]. Regarding SSL, our concern is that despite pretraining with state-of-the-art SSL techniques, the model still required large downstream datasets in the fine-tuning phase. For instance, even after pretraining with SSL, action recognition tasks required fine-tuning on large-scale Kinetics-400 for several epochs to achieve remarkable performance [329]. As for image-text retrieval, it is crucial to carefully design the text prompts to effectively bridge an emotion recognition model with a pretrained image-text model. Further research is needed to determine how pretraining techniques can substantially enhance emotion recognition.

2) *Comprehensive and Robust Body Landmarks*: The effectiveness of current 2-D pose estimation methods has led to their widespread usage in downstream tasks such as action recognition [97, 294]. However, for emotion recognition, these methods face significant limitations. First, body landmarks used in pose estimation are insufficient for the analysis of subtle emotions, as they do not account for all relevant movements. For instance, many methods only provide one landmark on the chest, and this limitation prevents skeleton-based emotion recognition algorithms from leveraging information regarding chest expansion or contraction, which is a common indicator of a person's confidence. Similarly, a person can express emotions through finger movements, which are not captured by most pose-estimation algorithms due to being too fine-grained. Developing an emotion-specific pose estimation method would require constructing a new, large-scale annotated dataset with additional landmarks, which can be a time-consuming and costly process. Second, 2-D pose estimation results can be noisy due to the jitter errors [330]. Although these errors may have a minimal effect on metrics of pose estimation benchmarks, they can significantly impact understanding of subtle bodily expressions, which demands substantially higher precision of landmark locations [51]. Given that pose estimation serves as the starting point for skeleton-based BEEU methods, any errors in human pose can have a ripple effect on a final emotion prediction.

3) *Accurate 3-D Pose or Mesh In-the-Wild*: The integration of 3-D pose information has the potential to substantially enhance BEEU algorithms because the extra dimension allows for a better understanding of movements in 3-D space. In other words, BEEU algorithms could depend on precise 3-D pose or mesh as input which should enhance its overall accuracy. Whereas accurate 3-D poses can be obtained through MoCap systems for laboratory-collected data, obtaining accurate 3-D poses for in-the-wild data is challenging. It is impossible to set up MoCap to capture the 3-D pose of a person because the area of movement is often too large for the placement of MoCap cameras. Furthermore, existing 3-D pose or mesh estimation approaches have poor performance when applied to in-the-wild images. The difficulty of collecting 3-D annotations for in-the-wild images has resulted in a lack of large-scale, high-quality 3-D pose datasets. Current 3-D pose models are heavily reliant on lab-collected 3-D datasets for training, which are

subject to a distinct domain shift from in-the-wild images. In a laboratory setting, the lighting and environment are fixed, and the appearance and posture of individuals are monotonous. Conversely, in-the-wild images exhibit significant variations in these factors, making it difficult for current models to generalize well to in-the-wild images.

B. Fundamental Statistical Modeling and Learning Methods

The field of emotion modeling has seen a shift from conventional computer vision and machine learning techniques (e.g., [46]) to deep learning-based methods (e.g., [331, 51]). For many other computer vision problems, deep learning has shown its power in substantially advancing the state of the art compared to other machine learning methods. However, some intrinsic limitations of deep learning continue to limit progress in the field of emotion modeling. In this context, we aim to highlight some fundamental data-driven AI capabilities that, if developed, could drive the field forward.

1) *Modeling a Complex Space with Scarce Data*: A person's emotions and behaviors are influenced by various factors such as personality, gender, age, race/ethnicity, cultural background, and situational context. Modeling this multidimensional, complex space requires a massive amount of data to sample it sufficiently and properly. Despite the potential to collect data from the Internet, manually annotating such a large dataset for emotion can be prohibitively expensive. This conundrum challenges machine learning researchers to find ways to learn meaningful information with limited data collection.

A relatively more specific challenge facing researchers developing datasets for emotion research is determining the appropriate amount of data needed to obtain meaningful results through machine learning. Emotion data, when collected from public sources, tends to be naturally imbalanced, with some emotion categories (e.g., happiness, sadness) having a much higher number of samples than others (e.g., yearning, pain) [51]. Unlike in typical object recognition where metadata (e.g., keywords, image file names) can aid in crawling a reasonably balanced dataset, we usually cannot determine or estimate the emotion label for a piece of visual data from just the available metadata. Making a balanced, representative dataset depends on crawling or collecting a very large dataset, in the hope of obtaining a sufficient number of samples for the less prevalent categories. Such a laborious process greatly increases the cost of data collection. A potential solution to this challenge is to use AI to guide a more efficient data collection process, so that limited annotation resources can be used to achieve maximum benefit for AI training.

2) *Explainability, Interpretability, and Trust*: The ability for AI systems to provide clear and interpretable explanations of their reasoning processes is particularly crucial in critical applications, such as healthcare, because end-users, who may not have a data science background, need to have confidence in the AI system's quantitative findings. There is ongoing research in this area within the broader field of AI, with some progress being made [332, 333, 334]. However, it is widely considered an open challenge.

The task of emotion understanding presents unique challenges in terms of explainability, interpretability, and trust. First, emotions are highly subjective, and AI models typically learn from a general population's responses based on data collected from numerous human subjects. The AI's ability to explain or interpret results, thus, is limited to a general perspective, which may naturally be considered unreasonable to a certain extent by virtually any particular individual, especially if their opinions on emotions frequently deviate from the norm.

Second, causal relationships are often more significant in emotion-related applications. For example, in determining why a sunset scene evokes positive emotions in many viewers, it is important to understand whether it is the sunset scenario, the orange hues, the gradient of colors, the horizontal composition, or some other properties that cause viewers to feel positive emotions. Without this knowledge, algorithms developed may behave unpredictably in many situations, acting as black boxes. Currently, methods for causal discovery are often not scalable to high-dimensional or complex data and are sensitive to sparse data. Therefore, there is a need to develop appropriate frameworks and methods for high-dimensional and complex causal discovery specifically tailored to the understanding of emotions.

Third, overlapping semantics among emotion labels add complexity to interpreting results. The BoLD dataset has shown, using video clip annotations, that several pairs of emotion labels are highly correlated [51]. Examples are pleasure and happiness (correlation = 0.57), happiness and excitement (0.40), sadness and suffering (0.39), annoyance and disapproval (0.37), sensitivity and sadness (0.37), and affection and happiness (0.35). Even in the dimensional VAD model, researchers have detected correlations between valence and dominance (0.359) and between arousal and dominance (0.356). Current data-driven approaches can often provide a probability score for each emotion label in the classification system. Although it remains common practice to sort these scores to determine the most likely emotion, it is not always clear what a mixture of scores represents. For example, what does it mean to have 80% happiness, 60% excitement, and 50% sadness? Is it reasonable to just assume the data point should be classified as happiness? Or, should it be considered a mixture of happiness, excitement, and sadness all at the same time, perhaps as representative of humans often complex emotional states? Or, could it be some partial combination of these categories? More fundamental statistical learning methods will likely be needed to address this issue in a principled way.

One potential strategy in automated emotion recognition is to uncover useful patterns from a large amount of data, with the aim of gaining a deeper understanding of emotions beyond simple classification. Such findings may either support or challenge existing psychological theories or inspire new hypotheses. However, black-box machine learning models can provide little insight into why a decision is made, which restricts our ability to gain knowledge through automated learning. As a result, there has been a recent and growing focus on developing interpretable machine learning.

There are two primary approaches to interpretable machine learning. The first is known as model-agnostic approaches [335, 336] and involves using relatively simple models such as decision trees and linear models to approximate locally the output of a black-box model like DNN. A fundamental issue with this approach is that the explanation is local, usually a neighborhood around every input point. This outcome raises the question of whether such limited explanations are truly useful. Because the power of explaining a phenomenon implies the capability to reveal underlying mechanisms coherently for a wide range of cases, a severe lack of generality in interpretation undermines this goal.

In the second approach, the emphasis is on developing models that are *inherently interpretable*. Classic statistical models, due to their simple structure, are often inherently interpretable. However, their accuracy is often significantly lower compared to top-performing black-box models. This drawback of classic models has motivated researchers to develop models with enhanced accuracy without losing interpretability. For example, Seo et al. [337] proposed the concept of cosupervision by DNN to train a mixture of linear models (MLM), aimed at filling the gap between transparent and black-box models. The idea is to treat a DNN model as an approximation to the optimal prediction function based on which augmented data are generated. Clustering methods are used to partition the feature space based on the augmented data, and a linear regression model (or logistic regression for classification) is fit in each region. Although MLMs have existed in various contexts, they have not been widely adopted for high-dimensional data because of the difficulties in generating a good partition. The authors overcame this challenge by exploiting DNN. They also developed methods to help interpret models either by visualization or simple description. Advances in the direction of developing accurate models that are directly interpretable are valuable for emotion recognition.

3) *Modeling Under Uncertainty*: Data available for modeling human emotions and behaviors is often suboptimal, leading to various challenges in accurate modeling. For instance, in the case of bodily expressed emotion recognition, the available data is often limited to partial body movements (e.g., only the upper body is visible in the video) and may include occlusions (i.e., certain body parts are blocked from view). In addition, the automated detection of human body landmarks in video frames is not always precise. An interesting research direction will be to establish accurate models based on incomplete and inaccurate data. Furthermore, it is important to quantify uncertainty throughout the machine learning process, given that such uncertainties are present at each step. This is necessary in order to effectively communicate results to users and help them make informed interpretations.

4) *Learning Paradigms for Ambiguous Concepts*: The utilization of traditional learning frameworks, such as DNN, SVM, and classification and regression trees (CART), may prove to be limited in tackling complex problems, such as modeling not-so-well-defined concepts like emotions and movements. Unlike more concrete objects, such as cars and apples, there may not always be a clear ground truth for expression of emotions in a video clip. The ambiguity of

establishing a ground truth can be influenced by subjective interpretations of annotators, leading to different viewpoints and no correct interpretation.

Although these traditional learning frameworks are optimized for well-defined concepts, their straightforward application to emotion recognition may lead to incorrect assumptions, such as assuming a majority annotator holds the ground truth. In the case of BEEU, the creation of a mid-layer of movement concepts between pixel-level information and high-level emotions brings forth another challenge. Many movement classes are qualitative in nature (e.g., dropping weight, rhythmicity, strong, light or smooth movement, and space harmony), further complicating the development of accurate models. The need for advanced learning paradigms that can effectively tackle the challenges of modeling complex, ambiguous concepts such as emotions and movements has become apparent.

5) *AI Fairness and Imbalanced Datasets*: In the collection of data regarding human emotions and behaviors, it is not uncommon for certain demographic or emotion/behavior groups to have smaller sample sizes compared to others. For instance, data about White people is often more abundant than data about Black people, and data about happiness and sadness is typically more abundant than data about esteem, fatigue, and annoyance. Similarly, data across cultural groups varies significantly. Despite being a current area of research in AI [338], tackling such inadequacies is particularly complex in emotion understanding due to its unique characteristics.

6) *Efficiency of AI Models*: For some applications, emotion recognition algorithms must be fit onto robots or mobile devices that are limited by their on-board computing hardware and battery power. Complex problems that require multiple AI algorithms and models to work in concert can be difficult to address in real-time without high-performance GPU/CPU computing hardware. Therefore, it is crucial to simplify the mathematical models while maintaining their accuracy.

Machine learning researchers have derived techniques to simplify the model, including channel pruning techniques. In an earlier attempt, Ye et al. [339] proposed a method for accelerating the computations of deep CNNs, which directly simplify the channel-to-channel computation graph without the need to perform the computationally difficult and not-always-useful task of making high-dimensional tensors of CNN structure sparse. There have been numerous recent studies on pruning neural networks [340, 341].

C. Fundamental AI Methods

These are some fundamental AI components for creating an effective human-AI interaction and collaboration system that can have a significant impact in relevant domains. Simply increasing the amount of data alone may not be sufficient to counterbalance the deficiencies in these fundamental areas of AI. Below we will explore some of these challenges in greater detail.

1) *Decision-Making Based on Complex Information*: In real-world applications, such as mental healthcare, we often need to incorporate multiple sources of information (e.g., nonverbal cues, verbal cues, health record information, observations over time), and different people may have different

sets of information inputs (e.g., some patients may not have observations over time whereas others may not have detailed health records). For example, when a patient with depression visits a clinic, a combination of behaviors, speech, and health record information can be used to make informed decisions. To effectively connect all relevant information, research is required to develop a comprehensive framework. This fundamental area of AI research has the potential to impact many AI applications.

2) *Integrative Computational Models and Simulations for Understanding Human Communication and Collaboration*: Current research primarily concentrates on analyzing individuals, such as deducing an individual's emotional state from their behavior. However, emotional expressions play a significant role in interpersonal communication and collaboration. The emotional expression of one individual can have a significant impact on the emotions and behavior of those around them. Thus, there is a need for the development of integrative computational models and simulations to investigate the emotional interactions among individuals. As the problem of understanding individual emotions remains unresolved, a comprehensive approach to study interpersonal emotional interactions must consider the uncertainties present in individual-level emotion recognition.

In addition, when robots are integrated into the interaction, the distinction between human-human and human-robot interactions (HRIs) must be weighed. The external design and behavior of robots can vary greatly and offer a much wider range of possibilities than humans, making it challenging to sample comprehensively the potential space of robot interactions. For example, a robot can take on various forms, such as a human, animal, or a unique entity with its own distinct personality. Similarly, robots can move in ways that go beyond human or animal-like motions, whereas research can be conducted with limitations on specific types of robots, the results may not be applicable to other forms of robots.

3) *Incorporation of Knowledge and Understanding of Cognitive Processes*: When humans interpret emotions through visual media, they rely on their accumulated knowledge and experience of the world around them. However, the same behavior can be interpreted differently depending on the context or situation. Data-driven emotion recognition approaches require vast amounts of training data to be effective, but the countless possible scenarios and contexts can make obtaining such data challenging. To address this issue, AI researchers have been exploring the development of common-sense knowledge representations [342]. Integrating these advances into an emotion recognition framework is an important area of research. In addition, cognitive scientists have gained valuable insights into human cognitive processes through experiments, and incorporating these findings into the design of a next-generation emotion recognition system could be key to its success.

The expression and interpretation of emotions by humans involve various levels and scales, ranging from basic physiological processes impacting a person's behavior to socio-cultural structures that shape their knowledge and actions. Currently, multilevel and multiscale analyses of emotions are

a rare occurrence in AI due to the complexity it entails.

4) *Prediction of Actions*: Most of the current research in the field focuses on emotion recognition using visual information. However, for certain human-AI interaction applications, it is necessary to not only recognize emotions based on past data but also proactively gather information in real-time and make predictions about future events. For example, in the event of a heated argument between two human workers, a robot may need to move closer to better understand the situation and based on changes in the behavior of the workers, it may need to predict any potential danger and take action to resolve the issue. This might entail alerting others or attempting to redirect the attention of involved parties. Research is required to map emotion recognition with appropriate actions, even while acknowledging the inherent uncertainty involved in the process of emotion understanding. Such process is more nuanced than typical scene understanding.

D. Demographics

Unlike many computer vision problems such as object detection, when it comes to emotion, we simply cannot ignore the effect of demographics. Emotional responses can vary greatly among different demographic groups, including gender, age, race, ethnicity, education, socioeconomic status, religion, marital status, sexual orientation, health and disability status, and psychiatric diagnosis. Existing machine learning-based recognition technologies are not equipped to effectively handle such a vast array of demographic factors. In the absence of sound methods for evaluation, our brains tend to resort to shortcuts that may not be dependable, in order to conserve energy and navigate difficult situations where solid judgment is not present. To fill the gap, we often rely on stereotypes, heuristics, experience, and limited understanding to gauge emotions in other demographic groups, which may be unreliable. However, when we design AI systems, such shortcuts are not acceptable as mistakes made by machines can have disproportionate negative consequences for individuals and society as a whole. Addressing the issue of demographics in automated emotion understanding will likely remain a persistent challenge in the field.

E. Disentangling Personality, Function, Emotion, and Style

A person's behavior, captured by imaging or movement sensors, is a combination of several elements, including personality, function, emotion, and style. Even if we can find solutions to the problems we have mentioned earlier, separating these elements so that emotional expression can be properly analyzed will remain challenging. For example, the same punching motion would convey different emotions in a volleyball game versus during an argument between two individuals. This single example highlights the need for AI to first understand the purpose of the movement. For the same function, with the exact sequence of movements, two persons with very different personalities and contexts would likely be expressing different emotions or at least different levels of the same emotion. Without knowing people's personality traits, it will be impossible to pinpoint their emotional state. There is

a need to advance technology to differentiate between these factors in movements.

While fine-tuning the learned model to a specific person is possible, it usually requires collecting a substantial amount of annotated data from that person, which may not be feasible in practical applications requiring personalization. Further research is necessary to develop methods for personalizing emotion-related models with minimal additional data collection.

F. Partitioning the Space of Emotion

Thus far, technology developers have mostly relied on psychological theories of emotion, including the various models we discussed. However, these models have limitations that make them not ideal for AI applications. For example, if a model used in an AI program has too many components, the program may struggle to differentiate among them. At the same time, if the model is too simple with too few components, the AI may not be able to fully grasp the human emotion for the intended application. The VAD model offers a solution to this issue, but it is not suitable for AI applications for which specific emotions need to be identified. A deeper understanding of the emotional spectrum in AI will lead to more effective applications.

In a recent study, Wortman and Wang articulated that the strongest models needed robust coverage, which meant defining the minimal core set of emotions from which all others could be derived [343]. Using techniques from natural language processing and statistical clustering, these researchers showed that a set of 15 discrete emotion categories could achieve maximum coverage. This finding applies across six major languages—Arabic, Chinese, English, French, Spanish, and Russian—they have tested. Categories were identified as *affable, affection, afraid, anger, apathetic, confused, happiness, honest, playful, rejected, sadness, spiteful, strange, surprised, and unhealthy*. A more refined model with 25 categories was also proposed, which included the addition of *accepted, despondent, enthusiasm, exuberant, fearless, frustration, loathed, reluctant, sarcastic, terrific, and yearning* and the removal of *rejected*. Through the analysis of two large-scale emotion recognition datasets, including BoLD, the researchers confirmed the superiority of their models compared to existing models [343].

G. Benchmarking Emotion Recognition

Effective benchmarking has been instrumental in driving advancements in various AI research areas. However, benchmarking for emotion recognition is a challenging task due to its unique nature and the obstacles discussed earlier. Below, we offer insights on how to establish meaningful benchmarks for the field of emotion recognition, with a specific emphasis on the relatively new area of BEEU, where benchmarking is currently lacking.

1) *Benchmark Task Types*: A suite of tasks should be devised, including basic tasks such as single-data-type recognition of emotion (based on video only, images only, audio only, skeleton only, human mesh only), as well as multimodal

recognition (a combination of video, audio, and text). Emotion localization, which involves determining the range of frames in a video that depicts a targeted label, as well as movement recognition or LMA recognition using video, skeleton, or human mesh should also be considered. Furthermore, tasks related to predicting emotion from movement coding and video, or based on interaction can be developed. With the rich data across various contexts, natural environments, or situations (e.g., celebration, disaster, learning), data mining tasks focused on social interactions, including the comparison between age groups or the impact of assistive animals on mood, can be explored. In addition, real-world use-case challenges targeting specific applications can be utilized to assess algorithms' broad applicability and robustness.

2) *Testing and Evaluation*: In a benchmarking competition, the performance of participating teams' algorithms or systems can be evaluated using various criteria. Along with standard prediction accuracy based on shared training and testing datasets and the extent of emotion coverage, a system's performance with limited training data and equitable accuracy across different demographic subgroups, such as gender or ethnicity, can also be considered. The competition host can supply training data of varying sizes and required metadata, allowing participating teams to focus on a specific evaluation criterion and compete against others using the same standard. This competition format promotes diverse scientific exploration among participating teams, and collectively, teams focusing on different standards broaden the scope of models being investigated, effectively fostering a form of free-style community collaboration.

3) *Verification and Validation*: To validate software packages developed by participating teams in the competition, winning teams should be required to deposit their packages on repositories such as GitHub. The competition host should provide guidelines for verification, including compatibility with common computing environments, comprehensive documentation, and clear feedback on the execution status and reasons for any unexpected termination. To maintain fairness, true labels for test cases should not be disclosed prior to the completion of a competition.

Software packages should undergo thorough verification and validation throughout the entire training and testing pipeline. The competition host should replicate the training and testing process provided by each winning team, and the results should be compared to the claimed results. To streamline the validation process, subsampling of test cases may be employed.

4) *Risk Management*: To gauge the robustness of winning teams' software packages, they should be asked to provide results from a set of robustness tests, although during the competition, the comparison standard should be based on a single, focused criterion within a relatively straightforward test framework. Specifically, the impact of variations in factors such as batch randomization, bootstrapped sampling of training images, and training data size should be numerically evaluated.

5) *Evaluation Metric*: As a person's emotional state does not fall exclusively into a single type, to provide a fair evaluation of algorithms developed by the competition participants,

emotions can be characterized by a distribution over a given set of types, allowing each dataset to be described by different (and multiple) types and users to select the set that works best for their methods. Both the ground truth and the output of emotion recognition algorithms are formatted as distributions over these types. Suppose there are a total of K emotion types denoted by e_1, e_2, \dots, e_K . Different from a typical classification problem, there is a more subtle relationship between these types. Each pair of emotion types has a specified distance (or similarity) instead of just being different. For example, the emotions "happy" and "sad" are farther apart than "happy" and "excited". As a result, when we compare two emotion distributions over these types, we want to account for the underlying distances between emotion types. These pairwise distances can be estimated from the data based on how two emotion types co-occur. We can then use Wasserstein distance [344] to compute the overall distance between the ground truth and the computer prediction. Other commonly used distances between distributions such as L_p norm or KL divergence cannot factor in the underlying distances between emotion types. Let the distance between e_i and e_j be $c_{i,j}$, $i, j = 1, \dots, K$. Consider two probability mass functions over $\{e_1, \dots, e_K\}$: $p = (p_1, \dots, p_K)$ and $q = (q_1, \dots, q_K)$. The Wasserstein distance is defined by an optimal transport problem [344]. Let $\mathbf{W} = (w_{i,j})_{i,j=1,\dots,K}$ be a non-negative matching matrix between p and q . The Wasserstein distance is:

$$\begin{aligned} D(p, q) &= \min_{\mathbf{W}} \sum_{i=1}^K \sum_{j=1}^K w_{i,j} c_{i,j} \\ w_{i,j} &\geq 0, i, j = 1, \dots, K \\ \sum_{i=1}^K w_{i,j} &= q_j, j = 1, \dots, K \\ \sum_{j=1}^K w_{i,j} &= p_i, i = 1, \dots, K. \end{aligned}$$

VI. BEYOND EMOTION: INTERACTION WITH OTHER DOMAINS

Emotion, as one of the core components of human-to-human communication, can play an essential role in an array of future technological advancements impacting different parts of society. In this section, we provide an overview of how visual emotional understanding can be connected with other research problems, domains, or application areas, including art and design (Sections VI-A and VI-B), mental health (Section VI-C), robotics, AI agents, autonomous vehicles, animation, and gaming (Section VI-D), information systems (Section VI-E), industrial safety (Section VI-F), and education (Section VI-G). Instead of attempting to provide exhaustive coverage, we aim to highlight key intersections. Because some areas are in their early stages of development, we provide only a brief discussion of their potential.

A. Emotion and Visual Art

Art often depicts human emotional expressions, conveys the artist's feelings, or evokes emotional responses in viewers.

Except for certain genres in visual art, e.g. Realism, achieving lifelikeness is usually not the primary goal. Dutch Post-Impressionist painter Vincent van Gogh wrote, “I want to paint what I feel, and feel what I paint.” Similarly, fine-art photographer Ansel Adams stated, “A great photograph is one that fully expresses what one feels, in the deepest sense, about what is being photographed.” It is evident that artists intentionally link visual elements in their works with emotions. However, the relationship between visual elements in art and the emotion they evoke is still largely an enigma.

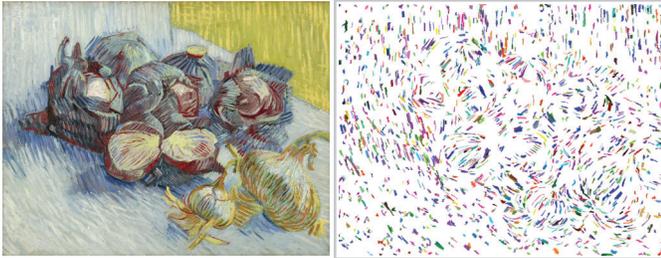


Fig. 18. Automatic brushstroke extraction for *Red Cabbages and Onions* (Paris, October–November 1887, oil on canvas, 50.2 cm × 64.3 cm) by Vincent van Gogh [345]. Painting image courtesy of the Van Gogh Museum, Amsterdam (Vincent van Gogh Foundation). The brushstroke map was provided by the James Z. Wang Research Group (The Pennsylvania State University).

Because visual artworks are almost always handcrafted by an artist and artists often develop unique styles, artworks are often abstract and difficult to analyze. In 2016, Lu et al. extended this research on evoked emotion in photographs [46] to paintings [47]. They acknowledged that using models developed for photographs on paintings would not be accurate due to different visual characteristics of the two types of images. To address this, they created an adaptive learning algorithm that leveraged labeled photographs and unlabeled paintings to infer the visual appeal of paintings.

To convey emotion effectively, artists often create and incorporate certain visual elements that are not commonly seen in real-world objects or scenes. An example is van Gogh’s highly rhythmic brushstroke style, which has been shown by computer vision researchers Li et al. to be one of the key characteristic differences between him and his contemporaries [345] (Fig. 18). In fact, van Gogh took piano lessons in the period 1883–1885 and in the middle of his painting career. He wrote to his younger brother Theo toward the latter part of his career in 1888, “.....this bloody mistral is a real nuisance for doing brushstrokes that hold together and intertwine well, with feeling, like a piece of music played with emotion.” The study by Li et al. [345] highlights the importance of designing algorithms specifically to answer the art-historical question at hand rather than using existing computer vision algorithms meant for analyzing real-world scenes.

Because artwork is the crystallization of the creativity and imagination of artists, studying artwork using computers and modern AI has the potential to reveal new perspectives on the connection between visual characteristics and emotion. Artists often incorporate exaggerated visual expressions such as carefully designed color palettes, tonal contrasts, brushstroke texture, and elegant curves. These features have inspired

computer scientists to create new algorithms for analyzing visual content. For instance, Yao et al. developed a color triplet analysis algorithm to predict the aesthetic quality of photographs, drawing inspiration from artists’ use of limited color palettes [346]. Li et al. created an algorithm for tonal adjustments based on the visual art concept of “Notan” which captures the dark and light arrangement of masses [347]. Motivated by the use of explicit and implicit triangles in artworks, He et al. and Zhou et al. developed algorithms to identify triangles in images, which can assist portrait and landscape photographers with composition [348, 349].

Current techniques for understanding emotion are not yet capable of analyzing certain aspects of emotion expressed in artwork, particularly at the level of composition or abstraction. For example, when emotions are conveyed through subtle interactions between people, the correspondence between low-level features we can extract and the emotions they represent cannot be easily established. American Impressionist painter Mary Cassatt’s work, for example, depicts the love bond between a mother and child through composition, pose, and brushstrokes, rather than through clear facial expressions. Similarly, American Modernist artist Georgia O’Keeffe used dramatic colors, elegant curves, and creative, abstract composition in her paintings of enlarged flowers and landscapes to convey feelings. She stated, “I had to create an equivalent for what I felt about what I was looking at – not copy it.” There is still much to be discovered by technology researchers in terms of the systematic connection between visual elements in abstract artwork and the emotions they convey.

B. Emotion and Design

Emotion plays a key role in product design, whether it be for physical or virtual products. Cognitive scientist D. A. Norman was a pioneer in the study of emotional design [350]. A successful design should evoke positive emotions in users/customers, such as excitement and a sense of pride and identity. In physical products, from a bottle of water to a house, designers carefully select visual elements, such as round corners, simple shapes, and elegant curves, to evoke positive emotions in customers. Similarly, designers of Websites, mobile apps, and other digital products and services use harmonious color schemes, simple and clean layouts, and emotion-provoking photographs to create a positive emotional impact on viewers or users.

By advancing evoked emotion prediction, future designers can be assisted by computers in multiple ways. First, computers can assess the evoked emotion of a draft design, based on models learned from large, annotated datasets. For example, a Website designer can ask the computer to rate a sample screenshot, identify areas for improvement, and provide advice on how to improve it. To develop this capability, however, researchers need to gain a better understanding of how demographics affect emotion. Certain design elements, e.g. color, may evoke different feelings in different cultures. A system trained with a general population may perform poorly for certain demographic groups. Second, computers can provide designers with design options that not only

meet customer needs but also evoke a specific emotion. For example, deep learning and generative adversarial networks (GANs) can already do the former task. If an additional emotion understanding component can be used to assess the design options generated and provide feedback to the system, the resulting designs can then evoke a desired emotion.

C. Emotion and Mental Health

Many mental health disorders can be considered disorders of emotion and emotion regulation [351]. This is because mental health disorders often entail extremes of chronic self-reported distress, sadness, anxiety, or lack of emotions such as flat affect and numbness as well as extremes in fluctuation of emotions [352]. For example, anxiety disorders are instances of being in an ongoing state of the fight or flight response, often viewing danger where none exists and therefore reacting as though the danger is constantly present and/or overreacting to ambiguous information. Emotion-related symptoms of anxiety disorders include panic attacks, fear, feelings of impending danger, agitation, excessive anxiety, feeling on edge, irritability, and restlessness [352]. Major depressive disorder has been conceptualized as a disorder of sustained negative affect, particularly sadness, and low levels of positive affect [353, 354]. Emotion-related symptoms of major depression include feeling sad or down most of the day nearly every day for at least two weeks and can also include an abundance of guilt, agitation, excessive crying, irritability, anxiety, apathy, hopelessness, loss of interest or pleasure in activities, mood swings, and feelings of restlessness. Similarly, bipolar disorder can include mood swings, elevated sadness, anger, anxiety, apathy, apprehension, euphoria, general discontent, guilt, hopelessness, loss of interest or pleasure, irritability, aggression, and agitation [352]. Extreme mood swings are also a prominent feature for some personality disorders such as borderline personality disorder, which also entails intense depressed mood, irritability or anxiety lasting a few hours to a few days, chronic feelings of emptiness, intense or uncontrollable anger, shame, and guilt [352]. Schizophrenia is associated with symptoms that are associated with mood. Positive symptoms can include delusions or paranoia, feeling anxious, agitated, or tense, and being jumpy or catatonic. Negative symptoms can include lack of interest or enthusiasm, lack of drive, and being emotionally flat [352]. Because extreme emotions are associated with these disorders researchers have examined ways to identify important distinctive features of them from videos.

Such studies have examined ways to use machine learning to code videos for nonverbal behaviors from facial expressions, body posture, gestures, voice analysis, and motoric functioning to diagnose mental health problems. In terms of facial expressions, studies have found that people with major depression, bipolar disorder, and schizophrenia demonstrated less facial expressivity compared to individuals without these disorders [355, 356, 357]. Depressed compared to non-depressed individuals also evidenced shorter durations and lower frequency of smiling behavior, less looking at an interviewer, and less eyebrow movement [358, 359, 360]. Such differences have been used to discriminate depressed

from nondepressed individuals [361, 362, 363, 364, 365, 366]. Studies have similarly diagnosed differential facial movement features of people with disorders such as autism spectrum disorder [367], posttraumatic stress disorder, and generalized anxiety disorder from healthy controls [368]. Similar to facial emotion, studies have used linguistic and voice emotion analysis to detect disorders such as depression, schizophrenia, bipolar disorder, posttraumatic stress disorder, and anxiety disorders [369, 370, 371].

As with facial actions, gestures and body movement have also been examined. In terms of gestures, those with depression showed more self-touching than those without depression [358, 360, 372]. Compared to schizophrenic individuals, those with depression tended to make fewer hand gestures [358]. Bipolar and depressed people also showed less gross motor activity than those without these disorders [360]. However, depressed and bipolar individuals showed more gross motor activity than people with schizophrenia [360]. At the same time, patients with schizophrenia demonstrated fewer hand gestures [373], fewer small and large head movements, and shorter duration of eye contact compared to those with depression [358, 360, 374]. Additional studies have detected attention deficit hyperactivity disorder from gestures and body movements [367]. Such differences have been used to diagnose mental health problems [375, 376]. See Table VII for more details about differentiating clinical disorders from healthy controls.

Gait, balance, and posture have also been used to identify mental health problems. For example, one meta-analysis summarized 33 studies of gait and balance in depression [474]. Depressed individuals had worse and more slumped posture [397, 395, 475, 393] and more postural instability and control [399]. In terms of gait, compared to healthy controls, those with depression took shorter strides, lifted their legs in an upward as opposed to a forward motion [391, 390], had more body sway, and walked more slowly, possibly to maintain their balance [393, 390, 401, 392]. These results are consistent with psychomotor retardation, a common symptom of depression. In terms of anxiety disorders, a study showed that these individuals walked more slowly, took shorter steps, and demonstrated problems with balance and mobility [463, 476]. Studies have also used gait to identify bipolar disorder [477], autism spectrum disorders [367], and attention deficit hyperactivity disorder [367].

In addition to mental health problems being disorders of emotional expression and experience, mental health problems can also be considered to be disorders of emotion recognition and understanding leading to social deficits. Understanding one's own and others' emotion has been termed theory of mind. Tests for theory of mind can include either identifying emotions by looking at images of faces with various emotional expressions (sometimes with parts of the faces obstructed), or watching a video of an interpersonal interaction and answering questions about various people's emotions and intentions. Having difficulties of theory of mind has been associated with depression [403, 404], social anxiety disorder [478], obsessive-compulsive disorder [479], schizophrenia [404, 480], bipolar disorder [404, 481], and autism spectrum disorders [473]. For

TABLE VII
DISTINCT EMOTION-RELATED NONVERBAL BEHAVIORS IN MENTAL HEALTH DISORDERS COMPARED TO HEALTHY CONTROLS.

Major Depressive Disorder	
<ul style="list-style-type: none"> • Reduced facial expressivity [355, 356] • Less eyebrow movements [379, 358, 380, 381] • Looking-down behaviors [382] • Reduced hand gestures [379, 381] • More self-touching [358, 383, 381, 360] • Reduced rate of speech [384] • Reduced speech [384] • Slower movements or abnormal postures [388, 372] • Reduced stride length and upward lifting motion of legs [390, 391] • Arm swing and vertical head movements while walking [393] • Slumped posture [394, 393, 395] • Balance difficulties during motor and cognitive tasks [398, 399, 393, 400, 401] • Impaired balance and lower gait velocity [390, 394, 393, 402] 	<ul style="list-style-type: none"> • Reduced variability of head movements [377, 378] • More nonspecific gaze patterns [382] • Less eye contact with another person [358] • Less smiling [379, 358, 380, 383, 381, 360, 359] • Slower voice [384] • More monotonic voice [385, 386, 369, 387] • Reduced pitch range [384] • Reduced gross motor activity [379, 389, 360] • Slower Gait speed [392, 390, 393] • Lateral upper body sway while walking [393] • Forward inclination of head and shoulders [396, 397] • Difficulty recognizing emotions [403, 404]
Bipolar Disorder	
<ul style="list-style-type: none"> • Reduced levels of facial expressivity [405] • Less gross motor activity [360] • More frequent and longer speech pauses when in depressive states [406, 409] • More postural sway [410] 	<ul style="list-style-type: none"> • Greater speech tonality [406, 407, 408] • Difficulty recognizing emotions [411, 412]
Schizophrenia	
<ul style="list-style-type: none"> • Reduced facial expressivity [413, 355, 356, 414] • Less smiling [358, 417, 418] • Reduced smiling eye gaze and head tilting associated with negative symptoms [419, 420] • Fewer hand gestures when speaking [373, 421, 419, 422, 420] • Less head nodding [373, 419] • Lower total time talking [423, 424, 371] • Longer speech pauses [423, 424, 371, 427] • Flat affect [371] • Balance difficulties and increased postural sway, [429, 430, 431, 432] • Slower velocity of walking and shorter strides [435] 	<ul style="list-style-type: none"> • Less upper facial movement expressing positive emotion [356, 415, 416] • Fewer gestures and poses [421] • Less head movement [374, 358] • Slower rate of speech [425, 426] • More pauses [371] • Forward head posture and lower spine curvature [428] • Difficulty walking in a straight line [433, 434] • Difficulty recognizing emotions [411, 412, 436]
Anxiety Disorders	
<ul style="list-style-type: none"> • Less eye contact [437, 438, 439] • Grimacing [437] • Strained face [442] • Smiled less, [438] • More and faster blinking [443, 446, 447, 448] • Small rapid head movements [446] • More physical movements indicative of nervousness (e.g., bouncing their knees, fidgety, repositing [449, 450, 451, 452, 453, 438] • Self touching [449] • Speech dysfluency [437] • Slow gait velocity associated with fear of falling [457, 458, 459] • Slower speed walking [463] • Enhanced recognition of anxiety [465, 466] 	<ul style="list-style-type: none"> • Instability of gaze direction [440] • Nonsymmetrical lip deformations [441] • Eyelid twitching [442] • More frequent and faster head movements [443, 444, 445] • Nodded less [438] • Made fewer gestures [438] • Speech errors [454] • More jittery voice [455, 456] • Balance dysfunction [460, 461, 462, 463, 464] • Shorter steps [463]
Posttraumatic Stress Disorder	
<ul style="list-style-type: none"> • Monotonous slower flatter speech [369, 467, 468, 469] • More anger, aggression, hostility, less joy [471, 470] 	<ul style="list-style-type: none"> • Reduced facial emotion [470]
Autism Spectrum Disorder	
<ul style="list-style-type: none"> • Distinctions in gait [472] 	<ul style="list-style-type: none"> • Difficulty recognizing emotions [473]

example, both schizophrenic and bipolar patients showed emotional reactivity that was discordant to emotional videos [357]. Studies have also examined videotapes of facial emotional reactions to emotionally evocative videos as a means of diagnosing mental health problems. For example, using this technique, those with autism spectrum disorders demonstrated impairment in their ability to recognize emotions from body gestures [482, 483]. Thus, emotion regulation, emotional understanding, and emotional reactivity can be impaired in those with mental health problems. Such impairment, however, can be used to create systems to automatically detect the presence of these emotional disorders. See Table VII for more details.

D. Emotion and Robotics, AI Agents, Autonomous Vehicles, Animation, and Gaming

A natural application domain for emotion understanding is robotics and AI. In science fiction films, robots and AI agents are often depicted as having a high level of EQ, such as R2-D2, T-800, and Wall-E. They are able to understand human emotion, effectively communicate their own emotional feelings, engage in emotional exchanges with other robots or humans, take appropriate actions in challenging situations or conflicts, and so on. The idea of empowering robots and AI with this level of EQ is widely seen as a desirable and ultimate goal, or a “Holy Grail.”

Some recent surveys studied the field of robotics and emotion [484, 485, 486], covering topics such as advanced sensors, latest modeling methods, and techniques for performing emotional actions. Research in the fields of HRI, human-machine interaction (HMI), and human-AI interaction is highly relevant. However, because BEEU is in its infancy and is considered a bottleneck technology, we have yet to see its applications in robotics.

If we can effectively model human emotions through both facial and bodily expressions, robots can work more effectively with human counterparts. Humans would be able to communicate with robots in a way similar to how emotions are used in human-to-human communication. For example, when human workers in a warehouse want to stop a fast-moving robot, they could wave their hands swiftly to signal distress. Similarly, pedestrians could wave their hands to signal to a self-driving vehicle on a highway that there is an accident ahead, and cars should slow down to avoid a collision. In such situations, traditional forms of communication such as speech and facial expressions may not effectively convey a sense of urgency.

Effective emotional communication can help us understand the intention of robots or AI. For instance, emotionless robots can be perceived as unfriendly or unsympathetic. In certain robotic or AI applications, such as companion robots or assistive robots, it is desirable to project a compassionate and supportive image to establish trust and cooperation between the device and humans interacting with it. Researchers have begun to investigate the relationship among robotics, personality, and motion [487].

In animated films, robots can display emotional behaviors, but these are often created by recording the movements of human actors through MoCap. That is, the animated characters mimic the movements of the actors behind the scenes. However, the capacity for computers to comprehend emotions akin to human perception could enable animated characters to use an emotion synthesis engine to autonomously generate authentic emotional behaviors. Advancements in computer graphics, virtual reality, and deep learning techniques including GANs, transformers, diffusion models, and contrastive learning have facilitated the creation of increasingly realistic and dynamic visual content. These technologies potentially enable the synthesis of complex and nuanced emotional displays.

Emotion understanding can substantially enhance the gaming experience by making games more emotionally responsive and immersive, as well as by providing personalized feedback to players. Game designers can make design decisions that enhance a player's experience based on the player's frustration level. If players are feeling sad, the game could offer them a story-based scenario that is more emotionally uplifting. By providing meaningful feedback, players are more likely to stay engaged with the game, improving their overall experience.

E. Emotion and Information Systems

Emotion understanding can play a pivotal role in advancing information systems. Currently, when searching for visual content in online collections, we primarily rely on keyword-based metadata. Whereas recent developments in deep learning

have enabled information systems to search using machine-generated annotations, these annotations are typically limited to identifying objects and relationships (e.g. a boy in a yellow shirt playing soccer).

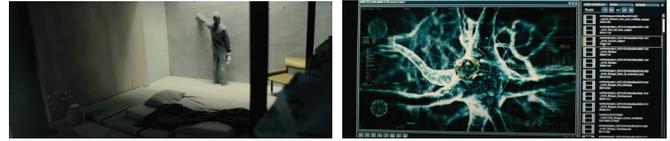


Fig. 19. Two scenes included in the computer-generated trailer for the sci-fi thriller “Morgan” [488]. The computer program considered these scenes most suited to give viewers a sense of horror or thrill. The copyright of these movie frames belongs to 20th Century Fox. They are used here for illustrating a scientific concept only.

IBM scientists demonstrated that computers with the ability to understand emotions could aid in sorting through a large amount of visual content and composing emotion-stimulating visual summaries [488]. In 1996, they created the first computer system for generating movie trailers. The trailer it produced for the 20th Century Fox sci-fi thriller “Morgan” was released as the official trailer. The system identified the top ten moments for inclusion in the trailer (Fig. 19). This work represents a significant milestone in understanding evoked emotions.

IBM's program is likely to be the starting point for a surge of emotion-based information systems. We can expect to see new applications such as evoked emotion assessment systems, emotion-based recommender systems, emotion-driven photo/video editing software, emotion-based media summarization/synopsizing services, and so on.

F. Emotion and Industrial Safety

Emotion understanding can be useful in promoting safety in workplaces such as factories and warehouses. It can provide early warnings of potential safety risks, such as worker fatigue or stress, allowing managers to take proactive measures to address the situation. Such capabilities can also provide personalized support and resources to workers who are experiencing emotional distress, improving overall emotional well-being and contributing to a safer work environment. The National Safety Council estimates that fatigue costs employers over \$130 billion annually in lost productivity, and over 70 million Americans have sleep disorders [489]. Existing research on fatigue detection typically involves specialized sensors or vision systems that monitor the face or eyes [490, 491, 492, 493, 494, 495, 496]. However, sensor-based approaches have limitations such as the need to wear them, size, cost, and reliability. Thus, there is a need to develop recognition systems that use body movement to enhance such systems [490, 497].

G. Emotion and Education

Emotion recognition technology can help create a more engaging and effective learning experience for online education. Many universities have been offering online courses for years, but the COVID-19 pandemic led to widespread adoption of online teaching using video conferencing platforms in 2020

and 2021. Even as in-person instruction has resumed, many educational institutions continue to conduct some of their teaching activities online. For example, instructors may be allowed to teach a portion of their classes online for pedagogical or emergency reasons or to hold office hours online. In a traditional classroom setting, instructors can gauge students' attentiveness and emotional states by observing their facial and bodily expressions. Such feedback can help instructors better understand and respond to students' needs, e.g., by adjusting the pace of instruction or covering alternative materials. However, in an online teaching environment, instructors often can only see the faces of a small number of students, which does not provide real-time feedback on the instruction. Potentially, if an online teaching platform could dynamically monitor the students and provide aggregated feedback (e.g., the percentage of students with high attentiveness, the overall mood of the class) to the instructor, the quality of online learning could be improved. To protect students' privacy, the monitoring process should only produce overall statistics.

VII. EMOTION AND SOCIETY: ETHICS

New technologies often bring new ethical concerns. In the field of emotion understanding, we have begun to witness the potential misuse of these technologies. In this section, we will discuss some of the general ethical issues surrounding the development and deployment of these technologies.

1) *Generalizability*: Because the emotion space is complex, it is important to recognize that there will always be outliers or unusual situations (e.g., an otherworldly scene, or an eccentric behavior) that may not be captured by our models. Without proper consideration of demographic differences and individual variations, these technologies may only provide a broad overview of the general population. To be truly beneficial, the system must be carefully tailored to specific needs of an individual. Likewise, diversity of representation in sample datasets is critical to ensure that algorithms emerging from them are inclusive.

2) *Verification of Accuracy or Performance*: It is important for researchers to keep in mind that there is *almost always* a lack of ground truth in emotion understanding. We have discussed the impact on data collection, modeling, and evaluation/benchmarking earlier (Sections III-B, V-B, and V-G). For AI models, it is imperative that both the output, design, and training processes are transparent and auditable. Black-box models can become uncontrollable if not properly monitored.

3) *Privacy – Data Collection*: The collection of human behavior data, including facial, body, and vocal information, raises privacy concerns. Research involving sensitive populations, such as patients in psychological clinics, must be conducted with utmost care. Further, almost all emotion-related annotations must be collected from humans. To protect human subjects, research protocols must be carefully designed to collect only necessary information, de-identify before distribution, and protect the data with proper access control and encryption. All protocols must be reviewed by an Institutional Review Board.

4) *Privacy – Use of Technology*: In today's automated world, people are losing their privacy to whoever controls data: Governments and companies are collecting data about where we are at any given moment; our financial transactions are followed and verified; companies are collecting data about our purchases, preferences, and social networks; most public places are constantly videotaped; and so on. People must sacrifice privacy to live a normal life because everything is computerized. We are being followed, and "Big brother" knows all about us. The only thing we can still keep to ourselves is our thoughts and emotions.

Once AI reaches high-accuracy automatic emotion recognition, our emotions will not be private anymore, and videos of our movements could be used against us by authorities or whoever will have videos of our movements. This situation could become very frightening. Moreover, if we want to hide our emotions, we will have to move in a way that will not reveal them, like using a "Poker face" to hide facial expressions. However, because specific movements not only express associated emotions but also enhance those emotions [115], moving in ways that flatten emotional expressions can also flatten the felt emotions, and living in such a way can lead to depression or other mental health problems.

5) *Synthesized Affective Behavior*: As much as its potential use in entertainment, success in emotion modeling could inevitably lead to even more lifelike deepfakes and similar abuses. As a society, instead of being fearful of the negative impact of new, beneficial technologies, we need to take on the challenge of detecting fakes, much as we recognize scammers, and mitigating the harm.

6) *Lower the Risks*: To mitigate the risks of misuse, proactive measures must be taken. It is essential that laws and regulations are established to keep pace with the rapid development of AEI technologies. As researchers, we have a responsibility to involve affected communities, particularly those that are traditionally marginalized such as minority groups, elderly individuals, and mental health patients, in design, development, and deployment processes to ensure that these individuals' perspectives and needs are recognized and valued.

7) *Performance Criteria*: To promote a responsible and ethical expansion of the field, it is crucial to establish reliable mechanisms for comparing the predictions of different algorithms and learning procedures. As discussed in Section V-G, a thorough evaluation of algorithms should not only consider accuracy and speed but also factors such as interpretability, demographic representation, context coverage, emotion space coverage, and personalization capabilities.

VIII. CONCLUSION

We provided an overview of the stimulating and exponentially growing field of visual emotion understanding. Adopting a multidisciplinary approach, we discussed the foundational principles guiding technological progress, reviewed recent innovations and system development, identified open challenges, and highlighted potential intersections with other fields. Our objective was to provide a comprehensive introduction to

this vast field sufficient to intrigue researchers across related IEEE subcommunities and to inspire continued research and development towards realizing the field's immense potential.

Given the multidisciplinary nature of this field, which encompasses multiple technical fields, psychology, and art, the barrier to entry can be considerable. Our aim is to provide researchers and developers with the essential knowledge required to tackle the numerous attractive open problems in the field. Interested readers are encouraged to delve deeper into the cited references for a more profound understanding of the topics discussed.

As active researchers in this domain, we strongly recommend that those interested in pursuing this research topic collaborate with others possessing complementary expertise. Although we anticipate the development and sharing of more large-scale datasets and continuous incremental progress, transformative solutions will not arise solely from straight application of data-driven approaches. Instead, sustained collaboration among researchers in computational, social and behavioral, and machine learning and statistical modeling fields will likely lead to lasting contributions to this intricate research field.

ACKNOWLEDGMENT

The authors would like to acknowledge the valuable contributions of several colleagues and advisees to this work. In particular, we would like to thank Jia Li for contributing valuable ideas and some writing related to explainable machine learning and benchmarking. They also extend appreciation to Hanjoo Kim, Amy LaViers, Xin Lu, Yu Luo, Yimu Pan, Nora Weibin Wang, Benjamin Wortman, Jianbo Ye, Sitao Zhang, and Lizhen Zhu for their research collaboration or discussions. They also thank Björn W. Schuller and Matti Pietikäinen for organizing this special issue on a timely and impactful topic. In addition, they would also like to express the gratitude to the anonymous reviewers for their valuable insights and constructive feedback, which contributed to the improvement of the manuscript. J. Z. Wang and C. Wu used the Extreme Science and Engineering Discovery Environment, which was supported by NSF under Grant ACI-1548562, and the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) Program supported by NSF under Grant OAC-2138259, Grant OAC-2138286, Grant OAC-2138307, Grant OAC-2137603, and Grant OAC-2138296. J. Z. Wang is grateful for the support and encouragement received from Adam Fineberg, Donald Geman, Robert M. Gray, Dennis A. Hejhal, Yelin Kim, Tatiana D. Korelsky, Edward H. Shortliffe, Juan P. Wachs, Gio Wiederhold, and Jie Yang throughout the years.

REFERENCES

- [1] F. Laricchia, "Smart speakers - Statistics & Facts," <https://www.statista.com/topics/4748/smart-speakers/#dossierKeyfigures>, 2022, accessed: 2022-07-6.
- [2] M. Krakovsky, "Artificial (emotional) intelligence," *Communications of the ACM*, vol. 61, no. 4, pp. 18–19, 2018.
- [3] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: A survey," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 43, no. 6, pp. 1815–1831, 2021.
- [4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Trans. Affective Computing*, vol. 13, no. 3, pp. 1195–1215, 2020.
- [5] M. Jampour and M. Javidi, "Multiview facial expression recognition, a survey," *IEEE Trans. Affective Computing*, vol. 13, no. 4, pp. 2086–2105, 2022.
- [6] Y. Liu, X. Zhang, Y. Li, J. Zhou, X. Li, and G. Zhao, "Graph-based facial affect analysis: A review," *IEEE Trans. Affective Computing*, 2022.
- [7] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5826–5846, 2022.
- [8] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Trans. Affective Computing*, vol. 9, no. 4, pp. 563–577, 2018.
- [9] Y. Li, J. Wei, Y. Liu, J. Kauttonen, and G. Zhao, "Deep learning for micro-expression recognition: A survey," *IEEE Trans. Affective Computing*, vol. 13, no. 4, pp. 2028–2046, 2022.
- [10] G. Brauwiers and F. Frasincar, "A survey on aspect-based sentiment classification," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–37, 2022.
- [11] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Trans. Affective Computing*, vol. 44, no. 9, p. 845–863, 2022.
- [12] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Trans. Affective Computing*, 2021.
- [13] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.
- [14] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: A survey," *IEEE Trans. Affective Computing*, 2020.
- [15] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Trans. Affective Computing*, 2021.
- [16] 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 Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6729–6751, 2021.
- [17] F. Noroozi, C. A. Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari, "Survey on emotional body gesture recognition," *IEEE Trans. Affective Computing*, vol. 12, no. 2, pp. 505–523, 2021.
- [18] M.-A. Mahfoudi, A. Meyer, T. Gaudin, A. Buendia, and S. Bouakaz, "Emotion expression in human body posture and movement: A survey on intelligible motion factors, quantification and validation," *IEEE Trans. Affective Computing*, 2022.
- [19] X. Li, Y. Zhang, P. Tiwari, D. Song, B. Hu, M. Yang, Z. Zhao, N. Kumar, and P. Martinen, "EEG based emotion recognition: A tutorial and review," *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–57, 2022.
- [20] S. Saganowski, B. Perz, A. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Trans. Affective Computing*, 2022.
- [21] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multimodal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.
- [22] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59–73, 2021.
- [23] G. Smith and J. Carette, "What lies beneath—a survey of affective theory use in computational models of emotion," *IEEE Trans. Affective Computing*, vol. 13, no. 4, pp. 1793–1812, 2022.
- [24] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, 2016.
- [25] 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.
- [26] Y. Wang, W. Song, W. Tao, A. Liotta, D. Yang, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang *et al.*, "A systematic review on affective computing: Emotion models, databases, and recent advances," *Information Fusion*, vol. 83–84, pp. 19–52, 2022.
- [27] A. Mehrabian, *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. The MIT Press, Cambridge, 1980.
- [28] D. Watson and A. Tellegen, "Toward a consensual structure of mood," *Psychological Bulletin*, vol. 98, no. 2, p. 219, 1985.
- [29] R. Thayer, *The Biopsychology of Mood and Arousal*. New York, NY: Oxford University Press, 1989.
- [30] R. J. Larsen and E. Diener, "Promises and problems with the circumplex model of emotion," in *Emotion*. Sage Publications, Inc, 1992.
- [31] B. Knutson, "Facial expressions of emotion influence interpersonal trait inferences," *J. Nonverbal Behavior*, vol. 20, no. 3, pp. 165–182, 1996.
- [32] C. Darwin and P. Prodger, *The Expression of the Emotions in Man and Animals*. Oxford University Press, USA, 1998.
- [33] W. James, *What is an Emotion?* Simon and Schuster, 2013.
- [34] P. M. Niedenthal, "Embodying emotion," *Science*, vol. 316, no. 5827, pp. 1002–1005, 2007.
- [35] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Proc. Nebraska Symposium on Motivation*. University of Nebraska Press, 1971.
- [36] L. F. Barrett, "Solving the emotion paradox: Categorization and the experience of emotion," *Personality and Social Psychology Review*, vol. 10, no. 1, pp. 20–46, 2006.
- [37] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect:

- An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.
- [38] J. A. Russell and J. M. Carroll, "On the bipolarity of positive and negative affect," *Psychological Bulletin*, vol. 125, no. 1, p. 3, 1999.
- [39] M. B. Arnold, *Emotion and Personality Vol. I. Psychological Aspects*. Columbia Univer. Press, 1960.
- [40] S. Schachter and J. Singer, "Cognitive, social, and physiological determinants of emotional state," *Psychological Review*, vol. 69, no. 5, p. 379, 1962.
- [41] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. the MIT Press, 1974.
- [42] J. A. Russell, "Core affect and the psychological construction of emotion," *Psychological Review*, vol. 110, no. 1, p. 145, 2003.
- [43] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior research methods*, vol. 37, no. 4, p. 626, 2005.
- [44] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 83–92.
- [45] P. J. Lang, M. M. Bradley, B. N. Cuthbert et al., *International Affective Picture System (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL, 2005.
- [46] X. Lu, P. Suryanarayan, R. B. Adams Jr, J. Li, M. G. Newman, and J. Z. Wang, "On shape and the computability of emotions," in *Proc. ACM Int. Conf. Multimedia*. ACM, 2012, pp. 229–238.
- [47] X. Lu, N. Sawant, M. G. Newman, R. B. Adams Jr, J. Z. Wang, and J. Li, "Identifying emotions aroused from paintings," in *Proc. European Conf. Computer Vision*, 2016, pp. 48–63.
- [48] X. Lu, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "An investigation into three visual characteristics of complex scenes that evoke human emotion," *Proc. Seventh Int. Conf. Affective Computing and Intelligent Interaction (ACII)*, 2017.
- [49] J. Ye, J. Li, M. G. Newman, R. B. Adams, and J. Z. Wang, "Probabilistic multigraph modeling for improving the quality of crowdsourced affective data," *IEEE Trans. Affective Computing*, vol. 10, no. 1, pp. 115–128, 2017.
- [50] H. Kim, X. Lu, M. Costa, B. Kandemir, R. B. Adams Jr, J. Li, J. Z. Wang, and M. G. Newman, "Development and validation of image stimuli for emotion elicitation (ISEE): A novel affective pictorial system with test-retest repeatability," *Psychiatry Research*, vol. 261, pp. 414–420, 2018.
- [51] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "ARBEE: Towards automated recognition of bodily expression of emotion in the wild," *Int. J. Computer Vision*, vol. 128, no. 1, pp. 1–25, 2020.
- [52] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Proc. European Conf. Computer Vision*, 2006, pp. 288–301.
- [53] A. J. Fridlund, *Human Facial Expression: An evolutionary view*. Academic press, 2014.
- [54] M. S. Yik, "Interpretation of faces: A cross-cultural study of a prediction from Fridlund's theory," *Cognition & Emotion*, vol. 13, no. 1, pp. 93–104, 1999.
- [55] N. H. Frijda and A. Tcherkassoff, "Facial expressions as modes of action readiness," in *The Psychology of Facial Expression*, J. A. Russell and J. M. Fernández-Dols, Eds. Editions de la Maison des Sciences de l'Homme, 1997.
- [56] R. J. Davidson, *Emotion and Affective Style: Hemispheric substrates*. SAGE Publications Sage CA: Los Angeles, CA, 1992.
- [57] N. E. Miller, "Analysis of the form of conflict reactions," *Psychological Bulletin*, vol. 34, no. 1, pp. 720–731, 1937.
- [58] E. Harmon-Jones, "Clarifying the emotive functions of asymmetrical frontal cortical activity," *Psychophysiology*, vol. 40, no. 6, pp. 838–848, 2003.
- [59] D. B. Centerbar, S. Schnall, G. L. Clore, and E. D. Garvin, "Affective incoherence: when affective concepts and embodied reactions clash," *J. Personality and Social Psychology*, vol. 94, no. 4, p. 560, 2008.
- [60] R. B. Adams, N. Ambady, C. N. Macrae, and R. E. Kleck, "Emotional expressions forecast approach-avoidance behavior," *Motivation and Emotion*, vol. 30, no. 2, pp. 177–186, 2006.
- [61] A. J. Nelson, R. B. Adams, M. T. Stevenson, M. Weisbuch, and M. I. Norton, "Approach-avoidance movement influences the decoding of anger and fear expressions," *Social Cognition*, vol. 31, no. 6, p. 745, 2013.
- [62] J. R. Averill, "A semantic atlas of emotional concepts," *JSAS: Catalog of Selected Documents in Psychology*, vol. 5, no. 330, pp. 1–64, 1975.
- [63] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, July 2017, pp. 1667–1675.
- [64] A. Wakabayashi, S. Baron-Cohen, S. Wheelwright, N. Goldenfeld, J. Delaney, D. Fine, R. Smith, and L. Weil, "Development of short forms of the empathy quotient (eq-short) and the systemizing quotient (sq-short)," *Personality and Individual Differences*, vol. 41, no. 5, pp. 929–940, 2006.
- [65] M. Kim and J. Leskovec, "Latent multi-group membership graph model," *arXiv preprint arXiv:1205.4546*, 2012.
- [66] L. Stappen, L. Schumann, B. Sertolli, A. Baird, B. Weigell, E. Cambria, and B. W. Schuller, "Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox," in *Proc. the 2nd Conf. Multimodal Sentiment Analysis Challenge*, 2021, pp. 75–82.
- [67] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2005. IEEE, 2005, pp. 381–385.
- [68] F. Zhou and F. Torre, "Canonical time warping for alignment of human behavior," *Advances in Neural Information Processing Systems*, vol. 22, 2009.
- [69] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Trans. Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.
- [70] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, 2022.
- [71] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proc. AAAI Conf. Artificial Intelligence*, 2016, pp. 308–314.
- [72] Y.-G. Jiang, B. Xu, and X. Xue, "Predicting emotions in user-generated videos," in *Proc. AAAI Conf. Artificial Intelligence*, 2014, pp. 73–79.
- [73] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Trans. Affective Computing*, vol. 9, no. 2, pp. 255–270, 2016.
- [74] T. Randhavane, U. Bhattacharya, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, "Identifying emotions from walking using affective and deep features," *arXiv preprint arXiv:1906.11884*, 2019.
- [75] X. Liu, H. Shi, H. Chen, Z. Yu, X. Li, and G. Zhao, "iMiGUE: An identity-free video dataset for micro-gesture understanding and emotion analysis," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 10 631–10 642.
- [76] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.
- [77] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int. J. Comput. Vision*, vol. 127, no. 6–7, p. 907–929, jun 2019.
- [78] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.
- [79] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "EMOTIC: Emotions in context dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2017, pp. 61–69.
- [80] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 03, pp. 34–41, 2012.
- [81] J. Lee, S. Kim, S. Kim, J. Park, and K. Sohn, "Context-aware emotion recognition networks," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2019, pp. 10 143–10 152.
- [82] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "DFEW: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 2881–2889.
- [83] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2022, pp. 20 922–20 931.
- [84] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "Samm: A spontaneous micro-facial movement dataset," *IEEE Trans. Affective Computing*, vol. 9, no. 1, pp. 116–129, 2016.
- [85] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "Cas (me) 2: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Trans. Affective Computing*, vol. 9, no. 4, pp. 424–436, 2017.
- [86] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, "YouTube movie reviews: Sentiment analysis in an audio-visual context," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.
- [87] A. B. 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 *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [88] P. Ekman and W. V. Friesen, *Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, Stanford University, Palo Alto, 1977.
- [89] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.
- [90] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Proc. Conf. Computer Vision and Pattern Recognition Workshop*. IEEE, 2006, pp. 149–149.
- [91] G. M. Jacob and B. Stenger, "Facial action unit detection with transformers," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 7680–7689.
- [92] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affective Computing*, vol. 10, no. 3, pp. 325–347, 2017.
- [93] R. Zhi, M. Liu, and D. Zhang, "A comprehensive survey on automatic facial action unit analysis," *The Visual Computer*, vol. 36, no. 5, pp. 1067–1093, 2020.
- [94] N. Wang, X. Gao, D. Tao, H. Yang, and X. Li, "Facial feature point detection: A comprehensive survey," *Neurocomputing*, vol. 275, pp. 50–65, 2018.
- [95] Y. Wu and Q. Ji, "Facial landmark detection: A literature survey," *Int. J. Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.
- [96] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. European Conf. Computer Vision*. Springer, 2014, pp. 740–755.

- [97] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.
- [98] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. European Conf. Computer Vision*, 2018, pp. 466–481.
- [99] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression," in *Proc. European Conf. Computer Vision*, 2018, pp. 529–545.
- [100] C. Wu, Y. Chen, J. Luo, C.-C. Su, A. Dawane, B. Hanzra, Z. Deng, B. Liu, J. Z. Wang, and C.-h. Kuo, "MEBOW: Monocular estimation of body orientation in the wild," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 3451–3461.
- [101] G. Moon, J. Y. Chang, and K. M. Lee, "Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2019, pp. 10 133–10 142.
- [102] C. Wu, Y. Li, X. Tang, and J. Z. Wang, "MUG: Multi-human graph network for 3D mesh reconstruction from 2D pose," *arXiv preprint arXiv:2205.12583*, 2022.
- [103] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (TOG)*, vol. 34, no. 6, p. 248, 2015.
- [104] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black, "Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image," in *Proc. European Conf. Computer Vision*, 2016, pp. 561–578.
- [105] F. Ahmed, A. H. Bari, and M. L. Gavrilova, "Emotion recognition from body movement," *IEEE Access*, vol. 8, pp. 11 761–11 781, 2019.
- [106] R. P. Tsachor and T. Shafir, "How shall I count the ways? A method for quantifying the qualitative aspects of unscripted movement with Laban Movement Analysis," *Frontiers in Psychology*, vol. 10, p. 572, 2019.
- [107] Z. Witkower and J. L. Tracy, "Bodily communication of emotion: Evidence for extrafacial behavioral expressions and available coding systems," *Emotion Review*, vol. 11, no. 2, pp. 184–193, 2019.
- [108] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Trans. Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.
- [109] L. Ebdali Takaloo, K. F. Li, and K. Takano, "An overview of emotion recognition from body movement," in *Proc. Computational Intelligence in Security for Information Systems Conference*, 2022, pp. 105–117.
- [110] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 4, pp. 1027–1038, 2011.
- [111] C. L. Roether, L. Omlor, A. Christensen, and M. A. Giese, "Critical features for the perception of emotion from gait," *J. Vision*, vol. 9, no. 6, pp. 15–15, 2009.
- [112] I. Bartenieff and D. Lewis, *Body Movement: Coping with the environment*. Gordon and Breach Science Publishers, 1980.
- [113] K. Studd and L. Cox, *Everybody is a Body*. Dog Ear Publishing, 2013.
- [114] A. Melzer, T. Shafir, and R. P. Tsachor, "How do we recognize emotion from movement? Specific motor components contribute to the recognition of each emotion," *Frontiers in Psychology*, vol. 10, p. 1389, 2019.
- [115] T. Shafir, R. P. Tsachor, and K. B. Welch, "Emotion regulation through movement: Unique sets of movement characteristics are associated with and enhance basic emotions," *Frontiers in Psychology*, vol. 6, p. 2030, 2016.
- [116] S. Saha, S. Datta, A. Konar, and R. Janarthanan, "A study on emotion recognition from body gestures using Kinect sensor," in *Proc. Int. Conf. Communication and Signal Processing*, 2014, pp. 056–060.
- [117] H. Zacharatos, C. Gatzoulis, P. Charalambous, and Y. Chrysanthou, "Emotion recognition from 3D motion capture data using deep CNNs," in *Proc. IEEE Conf. Games (CoG)*, 2021, pp. 1–5.
- [118] J. Shi, C. Liu, C. T. Ishi, and H. Ishiguro, "Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network," *Sensors*, vol. 21, no. 1, p. 205, 2020.
- [119] E. Ghaleb, A. Mertens, S. Asteriadis, and G. Weiss, "Skeleton-based explainable bodily expressed emotion recognition through graph convolutional networks," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FG 2021)*, 2021, pp. 1–8.
- [120] I. Ajili, M. Malle, and J.-Y. Didier, "Human motions and emotions recognition inspired by LMA qualities," *The Visual Computer*, vol. 35, no. 10, pp. 1411–1426, 2019.
- [121] A. Aristidou, P. Charalambous, and Y. Chrysanthou, "Emotion analysis and classification: Understanding the performers' emotions using the lma entities," *Computer Graphics Forum*, vol. 34, no. 6, pp. 262–276, 2015.
- [122] S. Dewan, S. Agarwal, and N. Singh, "Laban movement analysis to classify emotions from motion," in *Proc. Int. Conf. Machine Vision (ICMV 2017)*, vol. 10696. SPIE, 2018, pp. 717–724.
- [123] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann, "Continuous body emotion recognition system during theater performances," *Computer Animation and Virtual Worlds*, vol. 27, no. 3–4, pp. 311–320, 2016.
- [124] S. Wang, J. Li, T. Cao, H. Wang, P. Tu, and Y. Li, "Dance emotion recognition based on Laban motion analysis using convolutional neural network and long short-term memory," *IEEE Access*, vol. 8, pp. 124 928–124 938, 2020.
- [125] H. Cui, C. Maguire, and A. LaViers, "Laban-inspired task-constrained variable motion generation on expressive aerial robots," *Robotics*, vol. 8, no. 2, p. 24, 2019.
- [126] J. Inthiam, E. Hayashi, W. Jitviriyai, and A. Mowshowitz, "Development of an emotional expression platform based on LMA-shape and interactive evolution computation," in *Proc. Int. Conf. Control, Automation and Robotics (ICCAR)*, 2018, pp. 11–16.
- [127] M. M. Gross, E. A. Crane, and B. L. Fredrickson, "Effort-shape and kinematic assessment of bodily expression of emotion during gait," *Human Movement Science*, vol. 31, no. 1, pp. 202–221, 2012.
- [128] B. Adrian, *Actor Training the Laban Way: An integrated approach to voice, speech, and movement*. Simon and Schuster, 2010.
- [129] R. P. Tsachor and T. Shafir, "A somatic movement approach to fostering emotional resiliency through Laban Movement Analysis," *Frontiers in Human Neuroscience*, vol. 11, p. 410, 2017.
- [130] C. Fernandes, *The Moving Researcher: Laban/Bartenieff movement analysis in performing arts education and creative arts therapies*. Jessica Kingsley Publishers, 2014.
- [131] L. Bishko, "Animation principles and Laban movement analysis: Movement frameworks for creating empathic character performances," in *Nonverbal Communication in Virtual Worlds: Understanding and Designing Expressive Characters*, T. J. Tanenbaum, M. S. El-Nasr, and M. Nixon, Eds. Pittsburgh, PA: ETC Press, 2014, ch. 11, pp. 177–203.
- [132] J. van Geest, R. Samaritter, and S. van Hooren, "Move and be moved: The effect of moving specific movement elements on the experience of happiness," *Frontiers in Psychology*, vol. 11, p. 3974, 2021.
- [133] J. Z. Wang and J. Li, "Learning-based linguistic indexing of pictures with 2–d mhms," in *Proc. ACM Int. Conf. Multimedia*, 2002, pp. 436–445.
- [134] Y. Chen, J. Li, and J. Z. Wang, *Machine Learning and Statistical Modeling Approaches to Image Retrieval*. Springer Science & Business Media, 2006, vol. 14.
- [135] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, p. 5, 2008.
- [136] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using EMOTIC dataset," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755–2766, 2019.
- [137] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using frege's principle," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 14 234–14 243.
- [138] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.
- [139] T. Bänziger and K. R. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus," in *Proc. Int. Conf. Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 476–487.
- [140] B. Mesquita, M. Boiger, and J. De Leersnyder, "The cultural construction of emotions," *Current Opinion in Psychology*, vol. 8, pp. 31–36, 2016.
- [141] A. P. Atkinson, W. H. Dittrich, A. J. Gemmell, and A. W. Young, "Emotion perception from dynamic and static body expressions in point-light and full-light displays," *Perception*, vol. 33, no. 6, pp. 717–746, 2004.
- [142] S. L. Hutson-Comeaux and J. R. Kelly, "Gender stereotypes of emotional reactions: How we judge an emotion as valid," *Sex Roles*, vol. 47, no. 1, pp. 1–10, 2002.
- [143] T. M. Chaplin, "Gender and emotion expression: A developmental contextual perspective," *Emotion Review*, vol. 7, no. 1, pp. 14–21, 2015.
- [144] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures," *Emotion*, vol. 18, no. 1, p. 75, 2018.
- [145] R. Jürgens, A. Grass, M. Drolet, and J. Fischer, "Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected," *J. Nonverbal Behavior*, vol. 39, no. 3, pp. 195–214, 2015.
- [146] D. Keltner, D. Sauter, J. Tracy, and A. Cowen, "Emotional expression: Advances in basic emotion theory," *J. Nonverbal Behavior*, vol. 43, no. 2, pp. 133–160, 2019.
- [147] E. T. Hetzler, "Actors and emotion in performance," *Studies in Theatre and Performance*, vol. 28, no. 1, pp. 59–78, 2007.
- [148] N. K. Semmer, L. Messerli, and F. Tschann, "Disentangling the components of surface acting in emotion work: Experiencing emotions may be as important as regulating them," *J. Applied Social Psychology*, vol. 46, no. 1, pp. 46–64, 2016.
- [149] H. A. Elflein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychological Bulletin*, vol. 128, no. 2, p. 203, 2002.
- [150] R. Ekman, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.
- [151] D. Matsumoto, "Cultural similarities and differences in display rules," *Motivation and Emotion*, vol. 14, no. 3, pp. 195–214, 1990.
- [152] R. B. Adams Jr, U. Hess, and R. E. Kleck, "The intersection of gender-related facial appearance and facial displays of emotion," *Emotion Review*, vol. 7, no. 1, pp. 5–13, 2015.
- [153] A. H. Fischer, P. M. Rodriguez Mosquera, A. E. Van Vianen, and A. S. Manstead, "Gender and culture differences in emotion," *Emotion*, vol. 4, no. 1, p. 87, 2004.
- [154] R. B. Adams Jr, D. N. Albohn, N. Hedgecoth, C. O. Garrido, and K. D. Adams, "Angry white faces: A contradiction of racial stereotypes and emotion-resembling appearance," *Affective Science*, vol. 3, no. 1, pp. 46–61, 2022.
- [155] D. Matsumoto, "American-japanese cultural differences in judgements of expression intensity and subjective experience," *Cognition & Emotion*, vol. 13, no. 2, pp. 201–218, 1999.
- [156] H. Park and S. Kitayama, "Perceiving through culture: The socialized attention hypothesis," *The Science of Social Vision*, vol. 7, pp. 75–89, 2010.
- [157] J. Aronoff, B. A. Woike, and L. M. Hyman, "Which are the stimuli in facial displays of anger and happiness? configurational bases of emotion recognition," *J. Personality and Social Psychology*, vol. 62, no. 6, p. 1050, 1992.

- [158] M. Bar, M. Neta, and H. Linz, "Very first impressions." *Emotion*, vol. 6, no. 2, p. 269, 2006.
- [159] A. Arya, S. DiPaola, and A. Parush, "Perceptually valid facial expressions for character-based applications," *Int. J. Computer Games Technology*, vol. 2009, 2009.
- [160] F. Van Overwalle, T. Drenth, and G. Marsman, "Spontaneous trait inferences: Are they linked to the actor or to the action?" *Personality and Social Psychology Bulletin*, vol. 25, no. 4, pp. 450–462, 1999.
- [161] C. P. Said, N. Sebe, and A. Todorov, "Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces." *Emotion*, vol. 9, no. 2, p. 260, 2009.
- [162] S. Hareli and U. Hess, "What emotional reactions can tell us about the nature of others: An appraisal perspective on person perception," *Cognition and Emotion*, vol. 24, no. 1, pp. 128–140, 2010.
- [163] A. A. Marsh, R. B. Adams Jr, and R. E. Kleck, "Why do fear and anger look the way they do? form and social function in facial expressions," *Personality and Social Psychology Bulletin*, vol. 31, no. 1, pp. 73–86, 2005.
- [164] R. Donovan, A. Johnson, A. deRoiste, and R. O'Reilly, "Quantifying the links between personality sub-traits and the basic emotions," in *Computational Science and Its Applications—ICCSA 2020: 20th International Conference, Cagliari, Italy, July 1–4, 2020, Proceedings, Part II 20*. Springer, 2020, pp. 521–537.
- [165] D. J. Hughes, I. K. Kratsiotis, K. Niven, and D. Holman, "Personality traits and emotion regulation: A targeted review and recommendations." *Emotion*, vol. 20, no. 1, p. 63, 2020.
- [166] T. A. Fossum and L. F. Barrett, "Distinguishing evaluation from description in the personality-emotion relationship," *Personality and Social Psychology Bulletin*, vol. 26, no. 6, pp. 669–678, 2000.
- [167] J. M. Digman, "Personality structure: Emergence of the five-factor model," *Annual Review of Psychology*, vol. 41, no. 1, pp. 417–440, 1990.
- [168] L. R. Goldberg, "The structure of phenotypic personality traits." *American Psychologist*, vol. 48, no. 1, p. 26, 1993.
- [169] R. B. Adams Jr, A. J. Nelson, J. A. Soto, U. Hess, and R. E. Kleck, "Emotion in the neutral face: A mechanism for impression formation?" *Cognition & Emotion*, vol. 26, no. 3, pp. 431–441, 2012.
- [170] D. N. Albohn and R. B. Adams Jr, "The expressive triad: Structure, color, and texture similarity of emotion expressions predict impressions of neutral faces," *Frontiers in Psychology*, vol. 12, p. 612923, 2021.
- [171] H. S. Friedman, R. E. Riggio, and D. O. Segall, "Personality and the enactment of emotion," *J. Nonverbal Behavior*, vol. 5, no. 1, pp. 35–48, 1980.
- [172] R. Reisenzein, A. Hildebrandt, and H. Weber, "Personality and emotion," in *The Cambridge Handbook of Personality Psychology*, P. J. Corr and G. Matthews, Eds. Cambridge University Press, 2020, pp. 81–99.
- [173] D. Azucar, D. Marengo, and M. Settanni, "Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis," *Personality and Individual Differences*, vol. 124, pp. 150–159, 2018.
- [174] K. Lewin, *The Conceptual Representation and the Measurement of Psychological Forces*. Duke University Press, 1938.
- [175] C. S. Carver and M. F. Scheier, "Control theory: A useful conceptual framework for personality—social, clinical, and health psychology," *Psychological Bulletin*, vol. 92, no. 1, p. 111, 1982.
- [176] E. T. Higgins, "Self-discrepancy theory: What patterns of self-beliefs cause people to suffer?" in *Advances in Experimental Social Psychology*. Elsevier, 1989, vol. 22, pp. 93–136.
- [177] R. J. Davidson, "Affective style and affective disorders: Perspectives from affective neuroscience," *Cognition & Emotion*, vol. 12, no. 3, pp. 307–330, 1998.
- [178] J. A. Gray, *The Neuropsychology of Emotion and Personality*. Oxford University Press, 1987.
- [179] C. S. Carver and T. L. White, "Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales," *J. Personality and Social Psychology*, vol. 67, no. 2, p. 319, 1994.
- [180] R. J. Davidson and K. Hugdahl, *The Asymmetrical Brain*. MIT Press Cambridge, MA, 2003.
- [181] E. Harmon-Jones and J. J. Allen, "Anger and frontal brain activity: EEG asymmetry consistent with approach motivation despite negative affective valence," *J. personality and social psychology*, vol. 74, no. 5, p. 1310, 1998.
- [182] D. Joshi, R. Datta, E. Fedorovskaya, Q.-T. Luong, J. Z. Wang, J. Li, and J. Luo, "Aesthetics and emotions in images," *IEEE Signal Processing Magazine*, vol. 28, no. 5, pp. 94–115, 2011.
- [183] V. Yanulevskaya, J. C. van Gemert, K. Roth, A.-K. Herbold, N. Sebe, and J.-M. Geusebroek, "Emotional valence categorization using holistic image features," in *Proc. IEEE Int. Conf. Image Processing*, 2008, pp. 101–104.
- [184] L. Pang, S. Zhu, and C.-W. Ngo, "Deep multimodal learning for affective analysis and retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 2008–2020, 2015.
- [185] J. Yang, D. She, Y.-K. Lai, and M.-H. Yang, "Retrieving and classifying affective images via deep metric learning," in *Proc. AAAI Conf. Artificial Intelligence*, 2018, pp. 491–498.
- [186] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *Proc. IEEE Int. Conf. Image Processing*, 2016, pp. 634–638.
- [187] M. Muszynski, L. Tian, C. Lai, J. D. Moore, T. Kostoulas, P. Lombardo, T. Pun, and G. Chanel, "Recognizing induced emotions of movie audiences from multimodal information," *IEEE Trans. Affective Computing*, vol. 12, no. 1, pp. 36–52, 2021.
- [188] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, and X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 47–56.
- [189] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "Artemis: Affective language for visual art," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 11 569–11 579.
- [190] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 860–868.
- [191] Y. Bayeve, E. Dellandrea, C. Chamaret, and L. Chen, "Liris-accede: A video database for affective content analysis," *IEEE Trans. Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.
- [192] 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.
- [193] S. Zhao, H. Yao, Y. Gao, G. Ding, and T.-S. Chua, "Predicting personalized image emotion perceptions in social networks," *IEEE Trans. Affective Computing*, vol. 9, no. 4, pp. 526–540, 2018.
- [194] S. Zhao, G. Ding, Y. Gao, X. Zhao, Y. Tang, J. Han, H. Yao, and Q. Huang, "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE Trans. Affective Computing*, vol. 11, no. 4, pp. 574–587, 2020.
- [195] J. Song, K. Han, and S.-W. Kim, "'i have no text in my post': Using visual hints to model user emotions in social media," in *Proc. ACM Web Conference*, 2022, pp. 2888–2896.
- [196] X. Wang, J. Jia, J. Yin, and L. Cai, "Interpretable aesthetic features for affective image classification," in *Proc. IEEE Int. Conf. Image Processing*, 2013, pp. 3230–3234.
- [197] X. Liu, N. Li, and Y. Xia, "Affective image classification by jointly using interpretable art features and semantic annotations," *J. Visual Communication and Image Representation*, vol. 58, pp. 576–588, 2019.
- [198] A. Sartori, D. Culibrk, Y. Yan, and N. Sebe, "Who's afraid of itten: Using the art theory of color combination to analyze emotions in abstract paintings," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 311–320.
- [199] J. Yuan, S. Mcdonough, Q. You, and J. Luo, "Sentribute: image sentiment analysis from a mid-level perspective," in *Proc. Int. Workshop on Issues of Sentiment Discovery and Opinion Mining*, 2013, pp. 1–8.
- [200] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. ACM Int. Conf. Multimedia*, 2013, pp. 223–232.
- [201] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 159–168.
- [202] E. Acar, F. Hopfgartner, and S. Albayrak, "A comprehensive study on mid-level representation and ensemble learning for emotional analysis of video material," *Multimedia Tools and Applications*, vol. 76, no. 9, pp. 11 809–11 837, 2017.
- [203] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1106–1114, 2012.
- [204] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations*, 2015.
- [205] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [206] C. Xu, S. Cetintas, K. chih Lee, and L.-J. Li, "Visual sentiment prediction with deep convolutional neural networks," *arXiv preprint arXiv:1411.5731*, 2014.
- [207] H. Zhang and M. Xu, "Recognition of emotions in user-generated videos with kernelized features," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2824–2835, 2018.
- [208] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, "An end-to-end visual-audio attention network for emotion recognition in user-generated videos," in *Proc. AAAI Conf. Artificial Intelligence*, 2020, pp. 303–311.
- [209] X. Jin, P. Jing, J. Wu, J. Xu, and Y. Su, "Visual sentiment classification via low-rank regularization and label relaxation," *IEEE Trans. Cognitive and Developmental Systems*, 2021.
- [210] J. Wei, X. Yang, and Y. Dong, "User-generated video emotion recognition based on key frames," *Multimedia Tools and Applications*, vol. 80, no. 9, pp. 14 343–14 361, 2021.
- [211] A. Shukla, S. S. Gullapuram, H. Katti, M. Kankanhalli, S. Winkler, and R. Subramanian, "Recognition of advertisement emotions with application to computational advertising," *IEEE Trans. Affective Computing*, vol. 13, no. 2, pp. 781–792, 2022.
- [212] S. Deng, L. Wu, G. Shi, L. Xing, W. Hu, H. Zhang, and Y. Xiang, "Simple but powerful, a language-supervised method for image emotion classification," *IEEE Trans. Affective Computing*, 2022.
- [213] J. Pan, S. Wang, and L. Fang, "Representation learning through multimodal attention and time-sync comments for affective video content analysis," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 42–50.
- [214] C. Chen, Z. Wu, and Y.-G. Jiang, "Emotion in context: Deep semantic feature fusion for video emotion recognition," in *Proc. ACM Int. Conf. Multimedia*, 2016, pp. 127–131.
- [215] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artificial Intelligence*, 2015, pp. 381–388.
- [216] V. Campos, B. Jou, and X. Giro-i Nieto, "From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.
- [217] J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Proc. Int. Joint Conf. Artificial Intelligence*, 2017, pp. 3266–3272.

- [218] J. Yang, J. Li, L. Li, X. Wang, and X. Gao, "A circular-structured representation for visual emotion distribution learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 4237–4246.
- [219] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao, "Stimuli-aware visual emotion analysis," *IEEE Trans. Image Processing*, vol. 30, pp. 7432–7445, 2021.
- [220] Y. Liang, K. Maeda, T. Ogawa, and M. Haseyama, "Chain centre loss: A psychology inspired loss function for image sentiment analysis," *Neurocomputing*, vol. 495, pp. 118–128, 2022.
- [221] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 4489–4497.
- [222] Y. Ou, Z. Chen, and F. Wu, "Multimodal local-global attention network for affective video content analysis," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1901–1914, 2021.
- [223] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. AAAI Conf. Artificial Intelligence*, 2017, pp. 231–237.
- [224] K. Song, T. Yao, Q. Ling, and T. Mei, "Boosting image sentiment analysis with visual attention," *Neurocomputing*, vol. 312, pp. 218–228, 2018.
- [225] S. Fan, Z. Shen, M. Jiang, B. L. Koenig, J. Xu, M. S. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 7521–7531.
- [226] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer, "Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 192–201.
- [227] J. Yang, D. She, Y.-K. Lai, P. L. Rosin, and M.-H. Yang, "Weakly supervised coupled networks for visual sentiment analysis," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2018, pp. 7584–7592.
- [228] Z. Li, H. Lu, C. Zhao, L. Feng, G. Gu, and W. Chen, "Weakly supervised discriminate enhancement network for visual sentiment analysis," *Artificial Intelligence Review*, 2022.
- [229] X. Zhu, L. Li, W. Zhang, T. Rao, M. Xu, Q. Huang, and D. Xu, "Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition," in *Proc. Int. Joint Conf. Artificial Intelligence*, 2017, pp. 3595–3601.
- [230] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based convolutional neural network for image emotion classification," *Neurocomputing*, vol. 333, pp. 429–439, 2019.
- [231] T. Rao, X. Li, and M. Xu, "Learning multi-level deep representations for image emotion classification," *Neural Processing Letters*, vol. 51, no. 3, pp. 2043–2061, 2020.
- [232] J. Yang, X. Gao, L. Li, X. Wang, and J. Ding, "Solver: Scene-object interrelated visual emotion reasoning network," *IEEE Trans. Image Processing*, vol. 30, pp. 8686–8701, 2021.
- [233] H. Zhang and M. Xu, "Multiscale emotion representation learning for affective image recognition," *IEEE Trans. Multimedia*, 2022.
- [234] L. Xu, Z. Wang, B. Wu, and S. Lui, "Mdan: Multi-level dependent attention network for visual emotion analysis," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022, pp. 9479–9488.
- [235] J. Zhang, X. Liu, M. Chen, Q. Ye, and Z. Wang, "Image sentiment classification via multi-level sentiment region correlation analysis," *Neurocomputing*, vol. 469, pp. 221–233, 2022.
- [236] B. Xu, Y. Zheng, H. Ye, C. Wu, H. Wang, and G. Sun, "Video emotion recognition with concept selection," in *Proc. IEEE Int. Conf. Multimedia and Expo*, 2019, pp. 406–411.
- [237] J. J. Sun, T. Liu, and G. Prasad, "Gla in mediaeval 2018 emotional impact of movies task," *arXiv preprint arXiv:1911.12361*, 2019.
- [238] H. Cheng, Y. Tie, L. Qi, and C. Jin, "Context-aware based visual-audio feature fusion for emotion recognition," in *Proc. Int. Joint Conf. Neural Networks*, 2021, pp. 1–8.
- [239] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "DeepSentBank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.
- [240] X. Gu, L. Lu, S. Qiu, Q. Zou, and Z. Yang, "Sentiment key frame extraction in user-generated micro-videos via low-rank and sparse representation," *Neurocomputing*, vol. 410, pp. 441–453, 2020.
- [241] S. Ruan, K. Zhang, L. Wu, T. Xu, Q. Liu, and E. Chen, "Color enhanced cross correlation net for image sentiment analysis," *IEEE Trans. Multimedia*, 2021.
- [242] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
- [243] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.
- [244] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Trans. Knowledge and Data Engineering*, 2021.
- [245] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
- [246] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Trans. Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2010.
- [247] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.
- [248] P. Michel and R. El Kailouby, "Real time facial expression recognition in video using support vector machines," in *Proc. 5th Int. Conf. Multimodal Interfaces*, 2003, pp. 258–264.
- [249] 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 Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2008.
- [250] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.
- [251] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.
- [252] B. Fasel and J. Luetttin, "Automatic facial expression analysis: A survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.
- [253] C. Fabian Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "EmotioNet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 5562–5570.
- [254] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interaction*, 2017, pp. 553–560.
- [255] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proc. 18th ACM Int. Conf. Multimodal Interaction*, 2016, pp. 433–436.
- [256] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Proc. European Conf. Computer Vision*, 2016, pp. 499–515.
- [257] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Proc. IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 302–309.
- [258] X. Liu, B. Vijaya Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2017, pp. 20–29.
- [259] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutionary spatial-temporal networks," *IEEE Trans. Image Processing*, vol. 26, no. 9, pp. 4193–4203, 2017.
- [260] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, and Y. Zong, "Multi-cue fusion for emotion recognition in the wild," *Neurocomputing*, vol. 309, pp. 27–35, 2018.
- [261] X. Ouyang, S. Kawai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proc. 19th ACM Int. Conf. Multimodal Interaction*, 2017, pp. 577–582.
- [262] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3D convolutional networks," in *Proc. IEEE Int. Conf. Computer Vision Workshops*, 2017, pp. 1609–1618.
- [263] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proc. European Conf. Computer Vision*, 2018, pp. 222–237.
- [264] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906.
- [265] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2020, pp. 13984–13993.
- [266] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, and T. Mei, "Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 6248–6257.
- [267] Y. Chen and J. Joo, "Understanding and mitigating annotation bias in facial expression recognition," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 14980–14991.
- [268] D. Zeng, Z. Lin, X. Yan, Y. Liu, F. Wang, and B. Tang, "Face2exp: Combating data biases for facial expression recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2022, pp. 20291–20300.
- [269] H. Li, N. Wang, X. Ding, X. Yang, and X. Gao, "Adaptively learning facial expression representation via cf labels and distillation," *IEEE Trans. Image Processing*, vol. 30, pp. 2016–2028, 2021.
- [270] K. Wang, X. Peng, J. Yang, D. Meng, and Y. Qiao, "Region attention networks for pose and occlusion robust facial expression recognition," *IEEE Trans. Image Processing*, vol. 29, pp. 4057–4069, 2020.
- [271] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan, "Learning a facial expression embedding disentangled from identity," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 6759–6768.
- [272] C. Wang, S. Wang, and G. Liang, "Identity- and pose-robust facial expression recognition through adversarial feature learning," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 238–246.
- [273] D. Ruan, Y. Yan, S. Lai, Z. Chai, C. Shen, and H. Wang, "Feature decomposition and reconstruction learning for effective facial expression recognition," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, 2021, pp. 7660–7669.
- [274] A. H. Farzaneh and X. Qi, "Facial expression recognition in the wild via deep attentive center loss," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision*, 2021, pp. 2402–2411.
- [275] F. Xue, Q. Wang, and G. Guo, "Transfer: Learning relation-aware facial expression representations with transformers," in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 3601–3610.
- [276] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16

- words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learning Representations*, 2021.
- [277] A. V. Savchenko, L. V. Savchenko, and I. Makarov, “Classifying emotions and engagement in online learning based on a single facial expression recognition neural network,” *IEEE Trans. Affective Computing*, vol. 13, no. 4, pp. 2132–2143, 2022.
- [278] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Machine Learning*. PMLR, 2019, pp. 6105–6114.
- [279] D. Ruan, R. Mo, Y. Yan, S. Chen, J.-H. Xue, and H. Wang, “Adaptive deep disturbance-disentangled learning for facial expression recognition,” *Int. J. Computer Vision*, vol. 130, no. 2, pp. 455–477, 2022.
- [280] H. G. Wallbott, “Bodily expression of emotion,” *European J. Social Psychology*, vol. 28, no. 6, pp. 879–896, 1998.
- [281] H. K. Meeren, C. C. van Heijnsbergen, and B. de Gelder, “Rapid perceptual integration of facial expression and emotional body language,” *Proc. National Academy of Sciences of the United States of America*, vol. 102, no. 45, pp. 16518–16523, 2005.
- [282] B. De Gelder, “Towards the neurobiology of emotional body language,” *Nature Reviews Neuroscience*, vol. 7, no. 3, p. 242, 2006.
- [283] H. Aviezer, Y. Trope, and A. Todorov, “Body cues, not facial expressions, discriminate between intense positive and negative emotions,” *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.
- [284] N. L. Nelson and C. J. Mondloch, “Adults’ and children’s perception of facial expressions is influenced by body postures even for dynamic stimuli,” *Visual Cognition*, vol. 25, no. 4–6, pp. 563–574, 2017.
- [285] A. Karaaslan, B. Durmuş, and S. Amado, “Does body context affect facial emotion perception and eliminate emotional ambiguity without visual awareness?” *Visual Cognition*, vol. 28, no. 10, pp. 605–620, 2020.
- [286] H. Gunes and M. Piccardi, “Bi-modal emotion recognition from expressive face and body gestures,” *J. Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, 2007.
- [287] A. Kleinsmith, P. R. De Silva, and N. Bianchi-Berthouze, “Cross-cultural differences in recognizing affect from body posture,” *Interacting with Computers*, vol. 18, no. 6, pp. 1371–1389, 2006.
- [288] K. Schindler, L. Van Gool, and B. de Gelder, “Recognizing emotions expressed by body pose: A biologically inspired neural model,” *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008.
- [289] N. Dael, M. Mortillaro, and K. R. Scherer, “Emotion expression in body action and posture,” *Emotion*, vol. 12, no. 5, p. 1085, 2012.
- [290] B. Li, C. Zhu, S. Li, and T. Zhu, “Identifying emotions from non-contact gait information based on microsoft kinects,” *IEEE Trans. Affective Computing*, vol. 9, no. 4, pp. 585–591, 2016.
- [291] A. Crenn, R. A. Khan, A. Meyer, and S. Bouakaz, “Body expression recognition from animated 3D skeleton,” in *Proc. Int. Conf. 3D Imaging (IC3D)*, 2016, pp. 1–7.
- [292] U. Bhattacharya, C. Roncal, T. Mittal, R. Chandra, K. Kapsaskis, K. Gray, A. Bera, and D. Manocha, “Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping,” in *Proc. European Conf. Computer Vision*, 2020, pp. 145–163.
- [293] 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. Artificial Intelligence*, vol. 34, no. 02, 2020, pp. 1342–1350.
- [294] S. Yan, Y. Xiong, and D. Lin, “Spatial temporal graph convolutional networks for skeleton-based action recognition,” in *Proc. AAAI Conf. Artificial Intelligence*, 2018.
- [295] A. Banerjee, U. Bhattacharya, and A. Bera, “Learning unseen emotions from gestures via semantically-conditioned zero-shot perception with adversarial auto-encoders,” in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 3–10.
- [296] V. Narayanan, B. M. Manoghar, V. S. Dorbala, D. Manocha, and A. Bera, “Proxemo: Gait-based emotion learning and multi-view proxemic fusion for socially-aware robot navigation,” in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2020, pp. 8200–8207.
- [297] C. Hu, W. Sheng, B. Dong, and X. Li, “Tntc: two-stream network with transformer-based complementarity for gait-based emotion recognition,” in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 3229–3233.
- [298] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [299] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *Proc. European Conf. Computer Vision*, 2016, pp. 20–36.
- [300] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, “The kinetics human action video dataset,” *arXiv preprint arXiv:1705.06950*, 2017.
- [301] Y. Huang, H. Wen, L. Qing, R. Jin, and L. Xiao, “Emotion recognition based on body and context fusion in the wild,” in *Proc. IEEE/CVF Int. Conf. Computer Vision Workshops*, 2021, pp. 3609–3617.
- [302] P. P. Filntsis, N. Efthymiou, G. Potamianos, and P. Maragos, “Emotion understanding in videos through body, context, and visual-semantic embedding loss,” in *Proc. First Int. Workshop on Bodily Expressed Emotion Understanding, in conjunction with the European Computer Vision Conference*, 2020.
- [303] I. Pikoulis, P. P. Filntsis, and P. Maragos, “Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild,” in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition (FG 2021)*, 2021, pp. 01–08.
- [304] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Machine Learning*. PMLR, 2021, pp. 8748–8763.
- [305] S. Zhang, Y. Pan, and J. Z. Wang, “Learning emotion representations from verbal and nonverbal communication,” in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, June 2023.
- [306] H. Chen, H. Shi, X. Liu, X. Li, and G. Zhao, “SMG: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis,” *Int. J. Computer Vision*, pp. 1–21, 2023.
- [307] C. Wu, D. Davaasuren, T. Shafir, R. Tsachor, and J. Z. Wang, “Bodily expressed emotion understanding through integrating laban movement analysis,” *Patterns*, 2023, under revision for second-round review, arXiv preprint arXiv:2304.02187.
- [308] N. Le, K. Nguyen, A. Nguyen, and B. Le, “Global-local attention for emotion recognition,” *Neural Computing and Applications*, pp. 1–15, 2021.
- [309] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *Proc. Int. Conf. Machine Learning*. PMLR, 2021, pp. 5583–5594.
- [310] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423>
- [311] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “End-to-end multimodal emotion recognition using deep neural networks,” *IEEE J. Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
- [312] P. Antoniadis, I. Pikoulis, P. P. Filntsis, and P. Maragos, “An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild,” in *Proc. IEEE/CVF Int. Conf. Computer Vision*, 2021, pp. 3645–3651.
- [313] A. Shirian, S. Tripathi, and T. Guha, “Dynamic emotion modeling with learnable graphs and graph inception network,” *IEEE Trans. Multimedia*, vol. 24, pp. 780–790, 2021.
- [314] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, “Multi-task learning for multi-modal emotion recognition and sentiment analysis,” *arXiv preprint arXiv:1905.05812*, 2019.
- [315] W. Yu, H. Xu, Z. Yuan, and J. Wu, “Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis,” in *Proc. the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, 2021, pp. 10790–10797.
- [316] D. Jiang, R. Wei, H. Liu, J. Wen, G. Tu, L. Zheng, and E. Cambria, “A multitask learning framework for multimodal sentiment analysis,” in *Proc. Int. Conf. Data Mining Workshops (ICDMW)*. IEEE, 2021, pp. 151–157.
- [317] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, “Disentangled representation learning for multimodal emotion recognition,” in *Proc. ACM Int. Conf. on Multimedia*, 2022, pp. 1642–1651.
- [318] S. Zhang, C. Yin, and Z. Yin, “Multimodal sentiment recognition with multi-task learning,” *IEEE Trans. Emerging Topics in Computational Intelligence*, 2022.
- [319] K. Zhang, Y. Li, J. Wang, E. Cambria, and X. Li, “Real-time video emotion recognition based on reinforcement learning and domain knowledge,” *IEEE Trans. Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1034–1047, 2021.
- [320] T. Mittal, A. Bera, and D. Manocha, “Multimodal and context-aware emotion perception model with multiplicative fusion,” *IEEE MultiMedia*, vol. 28, no. 2, pp. 67–75, 2021.
- [321] E. Cambria, N. Howard, J. Hsu, and A. Hussain, “Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics,” in *IEEE symposium on computational intelligence for human-like intelligence (CIHLI)*. IEEE, 2013, pp. 108–117.
- [322] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proc. Conf. Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.
- [323] Z. Liu and Y. Shen, “Efficient low-rank multimodal fusion with modality-specific factors,” in *Proc. 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, 2018.
- [324] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, “Learning factorized multimodal representations,” in *Proc. Int. Conf. Learning Representations*, 2019. [Online]. Available: <https://openreview.net/forum?id=rygqqsA9KX>
- [325] Z. Sun, P. Sarna, W. Sethares, and Y. Liang, “Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis,” in *Proc. AAAI Conf. on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8992–8999.
- [326] D. Hazarika, R. Zimmermann, and S. Poria, “Misa: Modality-invariant and-specific representations for multimodal sentiment analysis,” in *Proc. ACM Int. Conf. on Multimedia*, 2020, pp. 1122–1131.
- [327] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.
- [328] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, “Masked autoencoders are scalable vision learners,” in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.
- [329] C. Feichtenhofer, H. Fan, Y. Li, and K. He, “Masked autoencoders as spatiotemporal learners,” *arXiv preprint arXiv:2005.09113*, 2022.
- [330] M. Ruggiero Ronchi and P. Perona, “Benchmarking and error diagnosis in multi-

- instance pose estimation,” in *Proc. IEEE Int. Conf. Computer Vision*, 2017, pp. 369–378.
- [331] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang, “Deep multi-patch aggregation network for image style, aesthetics, and quality estimation,” in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 990–998.
- [332] D. Shin, “The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai,” *Int. J. Human-Computer Studies*, vol. 146, p. 102551, 2021.
- [333] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic *et al.*, “Ai explainability 360: An extensible toolkit for understanding data and machine learning models,” *J. Mach. Learn. Res.*, vol. 21, no. 130, pp. 1–6, 2020.
- [334] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” in *Proc. IEEE 5th Int. Conf. Data Science and Advanced Analytics (DSAA)*, 2018, pp. 80–89.
- [335] M. T. Ribeiro, S. Singh, and C. Guestrin, “Model-agnostic interpretability of machine learning,” *arXiv preprint arXiv:1606.05386*, 2016.
- [336] —, ““Why should I trust you?” Explaining the predictions of any classifier,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [337] B. Seo, L. Lin, and J. Li, “Mixture of linear models co-supervised by deep neural networks,” *J. Computational and Graphical Statistics*, no. just-accepted, pp. 1–38, 2022.
- [338] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.
- [339] J. Ye, X. Lu, Z. Lin, and J. Z. Wang, “Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers,” in *Proc. Int. Conf. Learning Representation, arXiv preprint arXiv:1802.00124*, 2018.
- [340] J. Frankle and M. Carbin, “The lottery ticket hypothesis: Finding sparse, trainable neural networks,” *arXiv preprint arXiv:1803.03635*, 2018.
- [341] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden, and A. Peste, “Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks,” *J. Mach. Learn. Res.*, vol. 22, no. 241, pp. 1–124, 2021.
- [342] S. Ji, S. Pan, E. Cambria, P. Martinen, and S. Y. Philip, “A survey on knowledge graphs: Representation, acquisition, and applications,” *IEEE Trans. Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2021.
- [343] B. Wortman and J. Z. Wang, “HICEM: A high-coverage emotion model for artificial emotional intelligence,” *IEEE Trans. Affective Computing*, 2023, under revision for second-round review, arXiv preprint arXiv:2206.07593.
- [344] Y. Zhang, J. Z. Wang, and J. Li, “Parallel massive clustering of discrete distributions,” *ACM Trans. Multimedia Computing, Communications, and Applications (TOMM)*, vol. 11, no. 4, pp. 1–24, 2015.
- [345] J. Li, L. Yao, E. Hendriks, and J. Z. Wang, “Rhythmic brushstrokes distinguish van Gogh from his contemporaries: Findings via automated brushstroke extraction,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 6, pp. 1159–1176, 2012.
- [346] L. Yao, P. Suryanarayan, M. Qiao, J. Z. Wang, and J. Li, “OSCAR: On-site composition and aesthetics feedback through exemplars for photographers,” *Int. J. Computer Vision*, vol. 96, no. 3, pp. 353–383, 2012.
- [347] J. Li, L. Yao, and J. Z. Wang, “Photo composition feedback and enhancement,” in *Mobile Cloud Visual Media Computing*. Springer, 2015, pp. 113–144.
- [348] S. He, Z. Zhou, F. Farhat, and J. Z. Wang, “Discovering triangles in portraits for supporting photographic creation,” *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 496–508, 2017.
- [349] Z. Zhou, F. Farhat, and J. Z. Wang, “Detecting dominant vanishing points in natural scenes with application to composition-sensitive image retrieval,” *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2651–2665, 2017.
- [350] D. A. Norman, *Emotional Design: Why we love (or hate) everyday things*. Civitas Books, 2004.
- [351] G. Sheppes, G. Suri, and J. J. Gross, “Emotion regulation and psychopathology,” *Annual Review of Clinical Psychology*, vol. 11, pp. 379–405, 2015.
- [352] American Psychiatric Association, *Diagnostic and Statistical Manual of mental Disorders (DSM-5)*. American Psychiatric Association, 2013.
- [353] J. Joormann and C. H. Stanton, “Examining emotion regulation in depression: A review and future directions,” *Behaviour Research and Therapy*, vol. 86, pp. 35–49, 2016.
- [354] W. M. Vanderlind, Y. Millgram, A. R. Baskin-Sommers, M. S. Clark, and J. Joormann, “Understanding positive emotion deficits in depression: From emotion preferences to emotion regulation,” *Clinical Psychology Review*, vol. 76, p. 101826, 2020.
- [355] W. Gaebel and W. Wölwer, “Facial expression and emotional face recognition in schizophrenia and depression,” *European Archives of Psychiatry and Clinical Neuroscience*, vol. 242, no. 1, pp. 46–52, 1992.
- [356] —, “Facial expressivity in the course of schizophrenia and depression,” *European Archives of Psychiatry and Clinical Neuroscience*, vol. 254, no. 5, pp. 335–342, 2004.
- [357] G. Bersani, E. Polli, G. Valeriani, D. Zullo, C. Melcore, E. Capra, A. Quartini, P. Marino, A. Minichino, L. Bernabei *et al.*, “Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: A partially shared cognitive and social deficit of the two disorders,” *Neuropsychiatric Disease and Treatment*, vol. 9, p. 1137, 2013.
- [358] I. H. Jones and M. Pansa, “Some nonverbal aspects of depression and schizophrenia occurring during the interview,” *J. Nervous and Mental Disease*, vol. 167, no. 7, pp. 402–409, 1979.
- [359] A. Troisi and A. Moles, “Gender differences in depression: An ethological study of nonverbal behavior during interviews,” *J. Psychiatric Research*, vol. 33, no. 3, pp. 243–250, 1999.
- [360] C. Sobin and H. A. Sackeim, “Psychomotor symptoms of depression,” *American J. Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.
- [361] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, “Artificial intelligent system for automatic depression level analysis through visual and vocal expressions,” *IEEE Trans. Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2018.
- [362] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, “Automated depression diagnosis based on deep networks to encode facial appearance and dynamics,” *IEEE Trans. Affective Computing*, vol. 9, no. 4, pp. 578–584, 2018.
- [363] P. B. Kulkarni and M. M. Patil, “Clinical depression detection in adolescent by face,” in *Proc. Int. Conf. Smart City and Emerging Technology (ICSCET)*, 2018, pp. 1–4.
- [364] L. He, D. Jiang, and H. Sahli, “Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding,” *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1476–1486, 2019.
- [365] S. Song, L. Shen, and M. Valstar, “Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features,” in *Proc. IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 158–165.
- [366] S. A. Nasser, I. A. Hashim, and W. H. Ali, “A review on depression detection and diagnoses based on visual facial cues,” in *Proc. Int. Conf. Engineering Technology and its Applications (ICETA)*, 2020, pp. 35–40.
- [367] M. Mengi and D. Malhotra, “Artificial intelligence based techniques for the detection of socio-behavioral disorders: A systematic review,” *Archives of Computational Methods in Engineering*, pp. 1–45, 2021.
- [368] M. Gavrilescu and N. Vizireanu, “Predicting depression, anxiety, and stress levels from videos using the facial action coding system,” *Sensors*, vol. 19, no. 17, p. 3693, 2019.
- [369] D. M. Low, K. H. Bentley, and S. S. Ghosh, “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.
- [370] M. Morales, S. Scherer, and R. Levitan, “A cross-modal review of indicators for depression detection systems,” in *Proc. Fourth Workshop on Computational Linguistics and Clinical Psychology—From linguistic signal to clinical reality*, 2017, pp. 1–12.
- [371] A. Parola, A. Simonsen, V. Bliksted, and R. Fusaroli, “Voice patterns in schizophrenia: A systematic review and bayesian meta-analysis,” *Schizophrenia Research*, vol. 216, pp. 24–40, 2020.
- [372] D. Schrijvers, W. Hulstijn, and B. G. Sabbe, “Psychomotor symptoms in depression: A diagnostic, pathophysiological and therapeutic tool,” *J. Affective Disorders*, vol. 109, no. 1–2, pp. 1–20, 2008.
- [373] S. Anen, P. Roser, and M. Brüne, “Nonverbal behavior during clinical interviews: similarities and dissimilarities among schizophrenia, mania, and depression,” *The J. Nervous and Mental Disease*, vol. 200, no. 1, pp. 26–32, 2012.
- [374] P. Davison, C. Frith, P. Harrison-Read, and E. Johnstone, “Facial and other non-verbal communicative behaviour in chronic schizophrenia,” *Psychological Medicine*, vol. 26, no. 4, pp. 707–713, 1996.
- [375] A. Pampouchidou, M. Pediaditis, A. Maridakis, M. Awais, C.-M. Vazakopoulou, S. Sfakianakis, M. Tsiknakis, P. Simos, K. Marias, F. Yang *et al.*, “Quantitative comparison of motion history image variants for video-based depression assessment,” *EURASIP Journal on Image and Video Processing*, vol. 2017, no. 1, pp. 1–11, 2017.
- [376] A. Kacem, Z. Hammal, M. Daoudi, and J. Cohn, “Detecting depression severity by interpretable representations of motion dynamics,” in *Proc. IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 739–745.
- [377] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, “Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses,” *Image and Vision Computing*, vol. 32, no. 10, pp. 641–647, 2014.
- [378] S. Scherer, G. Stratou, and L.-P. Morency, “Audiovisual behavior descriptors for depression assessment,” in *Proc. 15th ACM on Int. Conf. Multimodal Interaction*, 2013, pp. 135–140.
- [379] M. JH Balsters, E. J Krahmer, M. GJ Swerts, and A. JJM Vingerhoets, “Verbal and nonverbal correlates for depression: A review,” *Current Psychiatry Reviews*, vol. 8, no. 3, pp. 227–234, 2012.
- [380] J. T. M. Schelde, “Major depression: Behavioral markers of depression and recovery,” *The J. Nervous and Mental Disease*, vol. 186, no. 3, pp. 133–140, 1998.
- [381] C. Segrin, “Social skills deficits associated with depression,” *Clinical Psychology Review*, vol. 20, no. 3, pp. 379–403, 2000.
- [382] T. Schelde and M. Hertz, “Ethology and psychotherapy,” *Ethology and sociobiology*, vol. 15, no. 5–6, pp. 383–392, 1994.
- [383] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, “Automatic behavior descriptors for psychological disorder analysis,” in *Proc. 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.
- [384] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, “A review of depression and suicide risk assessment using speech analysis,” *Speech Communication*, vol. 71, pp. 10–49, 2015.
- [385] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, “On the relative importance of vocal source, system, and prosody in human depression,” in *Proc. IEEE Int. Conf. Body Sensor Networks*, 2013, pp. 1–6.
- [386] G. Kiss and K. Vicsi, “Mono-and multi-lingual depression prediction based on speech processing,” *Int. J. Speech Technology*, vol. 20, no. 4, pp. 919–935, 2017.
- [387] T. F. Quatieri and N. Malyska, “Vocal-source biomarkers for depression: A link

to psychomotor activity," in *Thirteenth Annual Conf. of the International Speech Communication Association*, 2012, pp. 1059–1062.

[388] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin, "Psychomotor retardation in depression: biological underpinnings, measurement, and treatment," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 35, no. 2, pp. 395–409, 2011.

[389] G. Parker, D. Hadzi-Pavlovic, P. Boyce, K. Wilhelm, H. Brodaty, P. Mitchell, I. Hickie, and K. Eyers, "Classifying depression by mental state signs," *The British J. Psychiatry*, vol. 157, no. 1, pp. 55–65, 1990.

[390] M. R. Lemke, T. Wendorff, B. Mieth, K. Buhl, and M. Linnemann, "Spatiotemporal gait patterns during or ground locomotion in major depression compared with healthy controls," *J. Psychiatric Research*, vol. 34, no. 4-5, pp. 277–283, 2000.

[391] L. Sloman, M. Berridge, S. Homatidis, D. Hunter, and T. Duck, "Gait patterns of depressed patients and normal subjects," *The American J. Psychiatry*, vol. 139, pp. 94–97, 1982.

[392] J. M. Hausdorff, C.-K. Peng, A. L. Goldberger, and A. L. Stoll, "Gait unsteadiness and fall risk in two affective disorders: A preliminary study," *BMC Psychiatry*, vol. 4, no. 1, pp. 1–7, 2004.

[393] J. Michalak, N. F. Troje, J. Fischer, P. Vollmar, T. Heidenreich, and D. Schulte, "Embodiment of sadness and depression—gait patterns associated with dysphoric mood," *Psychosomatic Medicine*, vol. 71, no. 5, pp. 580–587, 2009.

[394] J. Michalak, J. Mischnat, and T. Teismann, "Sitting posture makes a difference—embodiment effects on depressive memory bias," *Clinical Psychology & Psychotherapy*, vol. 21, no. 6, pp. 519–524, 2014.

[395] C. Wilkes, R. Kydd, M. Sagar, and E. Broadbent, "Upright posture improves affect and fatigue in people with depressive symptoms," *J. Behavior Therapy and Experimental Psychiatry*, vol. 54, pp. 143–149, 2017.

[396] J. Z. Canales, J. T. Fiquer, R. N. Campos, M. G. Soeiro-de Souza, and R. A. Moreno, "Investigation of associations between recurrence of major depressive disorder and spinal posture alignment: A quantitative cross-sectional study," *Gait & Posture*, vol. 52, pp. 258–264, 2017.

[397] J. L. Rosario, M. S. B. Diógenes, R. Mattei, and J. R. Leite, "Differences and similarities in postural alterations caused by sadness and depression," *J. Bodywork and Movement Therapies*, vol. 18, no. 4, pp. 540–544, 2014.

[398] T. Deschamps, V. Thomas-Ollivier, A. Sauvaget, S. Bulteau, M. Fortes-Bourbousson, and H. Vachon, "Balance characteristics in patients with major depression after a two-month walking exercise program: A pilot study," *Gait & Posture*, vol. 42, no. 4, pp. 590–593, 2015.

[399] M. Doumas, C. Smolders, E. Brunfaut, F. Bouckaert, and R. T. Krampe, "Dual task performance of working memory and postural control in major depressive disorder," *Neuropsychology*, vol. 26, no. 1, p. 110, 2012.

[400] M. M. Nakano, T. S. Otonari, K. S. Takara, C. M. Carmo, and C. Tanaka, "Physical performance, balance, mobility, and muscle strength decline at different rates in elderly people," *J. Physical Therapy Science*, vol. 26, no. 4, pp. 583–586, 2014.

[401] S. Radovanović, M. Jovičić, N. P. Marić, and V. Kostić, "Gait characteristics in patients with major depression performing cognitive and motor tasks while walking," *Psychiatry Research*, vol. 217, no. 1-2, pp. 39–46, 2014.

[402] R. D. Sanders and P. M. Gillig, "Gait and its assessment in psychiatry," *Psychiatry (Edgmont)*, vol. 7, no. 7, pp. 38–43, 2010.

[403] E. Bora and M. Berk, "Theory of mind in major depressive disorder: A meta-analysis," *J. Affective Disorders*, vol. 191, pp. 49–55, 2016.

[404] T. van Neerven, D. J. Bos, and N. E. van Haren, "Deficiencies in theory of mind in patients with schizophrenia, bipolar disorder, and major depressive disorder: A systematic review of secondary literature," *Neuroscience & Biobehavioral Reviews*, vol. 120, pp. 249–261, 2021.

[405] M. A. Aghevli, J. J. Blanchard, and W. P. Horan, "The expression and experience of emotion in schizophrenia: A study of social interactions," *Psychiatry Research*, vol. 119, no. 3, pp. 261–270, 2003.

[406] A. Guidi, J. Schoentgen, G. Bertschy, C. Gentili, E. P. Scilingo, and N. Vanello, "Features of vocal frequency contour and speech rhythm in bipolar disorder," *Biomedical Signal Processing and Control*, vol. 37, pp. 23–31, 2017.

[407] A. Guidi, E. P. Scilingo, C. Gentili, G. Bertschy, L. Landini, and N. Vanello, "Analysis of running speech for the characterization of mood state in bipolar patients," in *Proc. AEIT Int. Annual Conf.*, 2015, pp. 1–6.

[408] J. Zhang, Z. Pan, C. Gui, T. Xue, Y. Lin, J. Zhu, and D. Cui, "Analysis on speech signal features of manic patients," *J. Psychiatric Research*, vol. 98, pp. 59–63, 2018.

[409] A. Maxhuni, A. Muñoz-Meléndez, V. Osmani, H. Perez, O. Mayora, and E. F. Morales, "Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients," *Pervasive and Mobile Computing*, vol. 31, pp. 50–66, 2016.

[410] A. R. Bolbecker, S. L. Hong, J. S. Kent, M. J. Klaunig, B. F. O'Donnell, and W. P. Hetrick, "Postural control in bipolar disorder: increased sway area and decreased dynamical complexity," *PLoS One*, vol. 6, no. 5, p. e19824, 2011.

[411] S. Baez, E. Herrera, L. Villarín, D. Theil, M. L. Gonzalez-Gadea, P. Gomez, M. Mosquera, D. Huepe, S. Strejilevich, N. S. Viglietta *et al.*, "Contextual social cognition impairments in schizophrenia and bipolar disorder," *PLoS One*, vol. 8, no. 3, p. e57664, 2013.

[412] G. Donohoe, A. Duignan, A. Hargreaves, D. W. Morris, E. Rose, D. Robertson, E. Cummings, S. Moore, M. Gill, and A. Corvin, "Social cognition in bipolar disorder versus schizophrenia: comparability in mental state decoding deficits," *Bipolar Disorders*, vol. 14, no. 7, pp. 743–748, 2012.

[413] H. Berenbaum and T. F. Oltmanns, "Emotional experience and expression in schizophrenia and depression," *J. Abnormal Psychology*, vol. 101, no. 1, p. 37, 1992.

[414] A. M. Kring and K. S. Earnst, "Stability of emotional responding in schizophrenia," *Behavior Therapy*, vol. 30, no. 3, pp. 373–388, 1999.

[415] G. Juckel and U. Polzer, "Fine analysis of abnormal facial expressions in chronic schizophrenic patients—a pilot study," *German J. Psychiatry*, vol. 1, pp. 6–9, 1998.

[416] R. Krause, E. Steimer, C. Säger-Alt, and G. Wagner, "Facial expression of schizophrenic patients and their interaction partners," *Psychiatry*, vol. 52, no. 1, pp. 1–12, 1989.

[417] E. Steimer-Krause, R. Krause, and G. Wagner, "Interaction regulations used by schizophrenic and psychosomatic patients: studies on facial behavior in dyadic interactions," *Psychiatry*, vol. 53, no. 3, pp. 209–228, 1990.

[418] A. Troisi, E. Pompili, L. Binello, and A. Sterpone, "Facial expressivity during the clinical interview as a predictor functional disability in schizophrenia. a pilot study," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 31, no. 2, pp. 475–481, 2007.

[419] M. Lavelle, P. G. Healey, and R. McCabe, "Is nonverbal communication disrupted in interactions involving patients with schizophrenia?" *Schizophrenia Bulletin*, vol. 39, no. 5, pp. 1150–1158, 2013.

[420] A. Troisi, "Ethological research in clinical psychiatry: the study of nonverbal behavior during interviews," *Neuroscience & Biobehavioral Reviews*, vol. 23, no. 7, pp. 905–913, 1999.

[421] M. Brüne, C. Sonntag, M. Abdel-Hamid, C. Lehmkämpfer, G. Juckel, and A. Troisi, "Nonverbal behavior during standardized interviews in patients with schizophrenia spectrum disorders," *The J. Nervous and Mental Disease*, vol. 196, no. 4, pp. 282–288, 2008.

[422] A. Troisi, G. Spalletta, and A. Pasini, "Non-verbal behaviour deficits in schizophrenia: An ethological study of drug-free patients," *Acta Psychiatrica Scandinavica*, vol. 97, no. 2, pp. 109–115, 1998.

[423] R. Kliper, S. Portuguese, and D. Weinsall, "Prosodic analysis of speech and the underlying mental state," in *Proc. Int. Symp. Pervasive Computing Paradigms for Mental Health*, 2015, pp. 52–62.

[424] R. Kliper, Y. Vaizman, D. Weinsall, and S. Portuguese, "Evidence for depression and schizophrenia in speech prosody," in *Proc. 3rd ICSA Tutorial and Research Workshop on Experimental Linguistics 2010*, 2010, pp. 85–88.

[425] C. Perlini, A. Marini, M. Garzitto, M. Isola, S. Cerruti, V. Marinelli, G. Rambaldelli, A. Ferro, L. Tomelleri, N. Dusi *et al.*, "Linguistic production and syntactic comprehension in schizophrenia and bipolar disorder," *Acta Psychiatrica Scandinavica*, vol. 126, no. 5, pp. 363–376, 2012.

[426] Y. Tahir, Z. Yang, D. Chakraborty, N. Thalmann, D. Thalmann, Y. Maniam, N. A. binte Abdul Rashid, B.-L. Tan, J. Lee Chee Keong, and J. Dauwels, "Non-verbal speech cues as objective measures for negative symptoms in patients with schizophrenia," *PLoS One*, vol. 14, no. 4, p. e0214314, 2019.

[427] V. Rapcan, S. D'Arcy, S. Yeap, N. Afzal, J. Thakore, and R. B. Reilly, "Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia," *Medical Engineering & Physics*, vol. 32, no. 9, pp. 1074–1079, 2010.

[428] V. B. Cristiano, M. F. Vieira Szortyka, M. I. Lobato, K. M. Ceresér, and P. Belmonte-de Abreu, "Postural changes in different stages of schizophrenia is associated with inflammation and pain: A cross-sectional observational study," *Int. J. Psychiatry in Clinical Practice*, vol. 21, no. 2, pp. 104–111, 2017.

[429] J. S. Kent, S. L. Hong, A. R. Bolbecker, M. J. Klaunig, J. K. Forsyth, B. F. O'Donnell, and W. P. Hetrick, "Motor deficits in schizophrenia quantified by nonlinear analysis of postural sway," *PLoS One*, vol. 7, no. 8, pp. 1–10, 2012.

[430] C. L. Marvel, B. L. Schwartz, and R. B. Rosse, "A quantitative measure of postural sway deficits in schizophrenia," *Schizophrenia Research*, vol. 68, no. 2-3, pp. 363–372, 2004.

[431] Y. Matsuura, H. Fujino, R. Hashimoto, Y. Yasuda, H. Yamamori, K. Ohi, M. Takeda, and O. Imura, "Standing postural instability in patients with schizophrenia: Relationships with psychiatric symptoms, anxiety, and the use of neuroleptic medications," *Gait & Posture*, vol. 41, no. 3, pp. 847–851, 2015.

[432] Y.-L. Teng, C.-L. Chen, S.-Z. Lou, W.-T. Wang, J.-Y. Wu, H.-I. Ma, and V. C.-H. Chen, "Postural stability of patients with schizophrenia during challenging sensory conditions: Implication of sensory integration for postural control," *PLoS One*, vol. 11, no. 6, p. e0158219, 2016.

[433] H. J. Jeon, M. J. Cho, S.-J. Cho, S.-U. Kim, S.-K. Park, J. S. Kwon, J. Y. Jeon, and B.-J. Hahn, "Quantitative analysis of ataxic gait in patients with schizophrenia: the influence of age and visual control," *Psychiatry Research*, vol. 152, no. 2-3, pp. 155–164, 2007.

[434] E. Lallart, R. Jouvent, F. R. Herrmann, F. Perez-Diaz, X. Lallart, O. Beauchet, and G. Allali, "Gait control and executive dysfunction in early schizophrenia," *J. Neural Transmission*, vol. 121, no. 4, pp. 443–450, 2014.

[435] A. Putzhammer, M. Perfahl, L. Pfeiff, and G. Hajak, "Gait disturbances in patients with schizophrenia and adaptation to treadmill walking," *Psychiatry and Clinical Neurosciences*, vol. 59, no. 3, pp. 303–310, 2005.

[436] A. Sparks, S. McDonald, B. Lino, M. O'Donnell, and M. J. Green, "Social cognition, empathy and functional outcome in schizophrenia," *Schizophrenia Research*, vol. 122, no. 1-3, pp. 172–178, 2010.

[437] B. O. Gilbert, "Physiological and nonverbal correlates of extraversion, neuroticism, and psychoticism during active and passive coping," *Personality and Individual Differences*, vol. 12, no. 12, pp. 1325–1331, 1991.

[438] A. Wenzel, J. Graff-Dolezal, M. Macho, and J. R. Brendle, "Communication and social skills in socially anxious and nonanxious individuals in the context of romantic relationships," *Behaviour Research and Therapy*, vol. 43, no. 4, pp. 505–519, 2005.

[439] A. N. Wiens, R. G. Harper, and J. D. Matarazzo, "Personality correlates of nonverbal interview behavior," *J. Clinical Psychology*, vol. 36, no. 1, pp. 205–215, 1980.

[440] G. Laretzaki, S. Plainis, I. Vrettos, A. Chrisoulakis, I. Pallikaris, and P. Bitsios,

- “Threat and trait anxiety affect stability of gaze fixation,” *Biological psychology*, vol. 86, no. 3, pp. 330–336, 2011.
- [441] D. Metaxas, S. Venkataraman, and C. Vogler, “Image-based stress recognition using a model-based dynamic face tracking system,” in *Proc. Int. Conf. Computational Science*, 2004, pp. 813–821.
- [442] M. Hamilton, “The assessment of anxiety states by rating,” *British J. Medical Psychology*, 1959.
- [443] D. F. Dinges, R. L. Rider, J. Dorrian, E. L. McGlinchey, N. L. Rogers, Z. Cizman, S. K. Goldenstein, C. Vogler, S. Venkataraman, and D. N. Metaxas, “Optical computer recognition of facial expressions associated with stress induced by performance demands,” *Aviation, Space, and Environmental Medicine*, vol. 76, no. 6, pp. B172–B182, 2005.
- [444] U. Hadar, T. J. Steiner, E. C. Grant, and F. Clifford Rose, “Head movement correlates of juncture and stress at sentence level,” *Language and Speech*, vol. 26, no. 2, pp. 117–129, 1983.
- [445] W. Liao, W. Zhang, Z. Zhu, and Q. Ji, “A real-time human stress monitoring system using dynamic bayesian network,” in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition-Workshops*, 2005, pp. 70–70.
- [446] G. Giannakakis, M. Pedititis, D. Manousos, E. Kazantzaki, F. Chiarugi, P. G. Simos, K. Marias, and M. Tsiknakis, “Stress and anxiety detection using facial cues from videos,” *Biomedical Signal Processing and Control*, vol. 31, pp. 89–101, 2017.
- [447] J. A. Harrigan, K. M. Harrigan, B. A. Sale, and R. Rosenthal, “Detecting anxiety and defensiveness from visual and auditory cues,” *J. Personality*, vol. 64, no. 3, pp. 675–709, 1996.
- [448] C. S. Harris, R. I. Thackray, and R. W. Shoenberger, “Blink rate as a function of induced muscular tension and manifest anxiety,” *Perceptual and Motor Skills*, vol. 22, no. 1, pp. 155–160, 1966.
- [449] P. Ekman and W. V. Friesen, “Detecting deception from the body or face,” *J. Personality and Social Psychology*, vol. 29, no. 3, p. 288, 1974.
- [450] E. A. Heerey and A. M. Krings, “Interpersonal consequences of social anxiety,” *J. Abnormal Psychology*, vol. 116, no. 1, p. 125, 2007.
- [451] A. P. Jurich and J. A. Jurich, “Correlations among nonverbal expressions of anxiety,” *Psychological Reports*, vol. 34, no. 1, pp. 199–204, 1974.
- [452] W. A. LeCompte, “The ecology of anxiety: Situational stress and rate of self-stimulation in turkey,” *J. Personality and Social Psychology*, vol. 40, no. 4, pp. 712–721, 1981.
- [453] T. Shechter, M. Asher, and I. M. Aderka, “Man vs. machine: A comparison of human and computer assessment of nonverbal behavior in social anxiety disorder,” *J. Anxiety Disorders*, vol. 89, p. 102587, 2022.
- [454] J. A. Harrigan, M. A. Larson, and C. J. Pflum, “The role of auditory cues in the detection of state anxiety 1,” *J. Applied Social Psychology*, vol. 24, no. 22, pp. 1965–1983, 1994.
- [455] T. Özseven, M. Dügenci, A. Doruk, and H. I. Kahraman, “Voice traces of anxiety: Acoustic parameters affected by anxiety disorder,” *Archives of Acoustics*, pp. 625–636, 2018.
- [456] V. Silber-Varod, H. Kreiner, R. Lovett, Y. Levi-Belz, and N. Amir, “Do social anxiety individuals hesitate more? the prosodic profile of hesitation disfluencies in social anxiety disorder individuals,” in *Proceedings of Speech Prosody*, vol. 2016, 2016, pp. 1211–1215.
- [457] M. F. Reelick, M. B. van Iersel, R. P. Kessels, and M. G. O. Rikkert, “The influence of fear of falling on gait and balance in older people,” *Age and Ageing*, vol. 38, no. 4, pp. 435–440, 2009.
- [458] J. P. Staab, C. D. Balaban, and J. M. Furman, “Threat assessment and locomotion: clinical applications of an integrated model of anxiety and postural control,” in *Seminars in Neurology*, vol. 33, no. 03. Thieme Medical Publishers, 2013, pp. 297–306.
- [459] D. Wynaden, J. Tohotoa, K. Heslop, and O. Al Omari, “Recognising falls risk in older adult mental health patients and acknowledging the difference from the general older adult population,” *Collegian*, vol. 23, no. 1, pp. 97–102, 2016.
- [460] C. D. Balaban, “Neural substrates linking balance control and anxiety,” *Physiology & Behavior*, vol. 77, no. 4–5, pp. 469–475, 2002.
- [461] O. Bart, Y. Bar-Haim, E. Weizman, M. Levin, A. Sadeh, and M. Mintz, “Balance treatment ameliorates anxiety and increases self-esteem in children with comorbid anxiety and balance disorder,” *Research in Developmental Disabilities*, vol. 30, no. 3, pp. 486–495, 2009.
- [462] B. Bolmont, P. Gangloff, A. Vouriot, and P. P. Perrin, “Mood states and anxiety influence abilities to maintain balance control in healthy human subjects,” *Neuroscience Letters*, vol. 329, no. 1, pp. 96–100, 2002.
- [463] R. Feldman, S. Schreiber, C. G. Pick, and E. Been, “Gait, balance, mobility and muscle strength in people with anxiety compared to healthy individuals,” *Human Movement Science*, vol. 67, p. 102513, 2019.
- [464] J.-P. Hainaut, G. Cailliet, F. G. Lestienne, and B. Bolmont, “The role of trait anxiety on static balance performance in control and anxiogenic situations,” *Gait & Posture*, vol. 33, no. 4, pp. 604–608, 2011.
- [465] P. Surcinelli, M. Codispoti, O. Montebanacci, N. Rossi, and B. Baldaro, “Facial emotion recognition in trait anxiety,” *J. Anxiety Disorders*, vol. 20, no. 1, pp. 110–117, 2006.
- [466] N. H. Zainal and M. G. Newman, “Worry amplifies theory-of-mind reasoning for negatively valenced social stimuli in generalized anxiety disorder,” *J. Affective Disorders*, vol. 227, pp. 824–833, 2018.
- [467] C. R. Marmar, A. D. Brown, M. Qian, E. Laska, C. Siegel, M. Li, D. Abu-Amara, A. Tsiartas, C. Richey, J. Smith *et al.*, “Speech-based markers for posttraumatic stress disorder in us veterans,” *Depression and Anxiety*, vol. 36, no. 7, pp. 607–616, 2019.
- [468] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, “Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews,” *IEEE Trans. Affective Computing*, vol. 7, no. 1, pp. 59–73, 2015.
- [469] R. Xu, G. Mei, G. Zhang, P. Gao, T. Judkins, M. Cannizzaro, and J. Li, “A voice-based automated system for PTSD screening and monitoring,” *Stud Health Technol Inform*, vol. 173, pp. 552–558, 2012.
- [470] G. Stratu, S. Scherer, J. Gratch, and L.-P. Morency, “Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences,” in *Proc. Humaine Association Conf. Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 147–152.
- [471] A. Kirsch and S. Brunnhuber, “Facial expression and experience of emotions in psychodynamic interviews with patients with ptsd in comparison to healthy subjects,” *Psychopathology*, vol. 40, no. 5, pp. 296–302, 2007.
- [472] E. Biffi, C. Costantini, S. B. Ceccarelli, A. Cesareo, G. M. Marzocchi, M. Nobile, M. Molteni, and A. Crippa, “Gait pattern and motor performance during discrete gait perturbation in children with autism spectrum disorders,” *Frontiers in Psychology*, vol. 9, p. 2530, 2018.
- [473] S. Baron-Cohen, “Theory of mind and autism: A review,” *International Review of Research in Mental Retardation*, vol. 23, pp. 169–184, 2000.
- [474] M. B. Murri, F. Triolo, A. Coni, C. Tacconi, E. Nerozzi, A. Escelsior, M. Respino, F. Neviani, M. Bertolotti, K. Bertakis *et al.*, “Instrumental assessment of balance and gait in depression: A systematic review,” *Psychiatry Research*, vol. 284, p. 112687, 2020.
- [475] J. Z. Canales, T. A. Cordás, J. T. Fiquer, A. F. Cavalcante, and R. A. Moreno, “Posture and body image in individuals with major depressive disorder: A controlled study,” *Brazilian J. Psychiatry*, vol. 32, pp. 375–380, 2010.
- [476] R. Feldman, S. Schreiber, C. Pick, and E. Been, “Gait, balance and posture in major mental illnesses: Depression, anxiety and schizophrenia,” *Austin Medical Sciences*, vol. 5, no. 1, pp. 1–6, 2020.
- [477] G. E. Kang, B. J. Mickey, M. G. McInnis, B. S. Krembs, and M. M. Gross, “Motor behavior characteristics in various phases of bipolar disorder revealed through biomechanical analysis: Quantitative measures of activity and energy variables during gait and sit-to-walk,” *Psychiatry Research*, vol. 269, pp. 93–101, 2018.
- [478] D. M. Hezel and R. J. McNally, “Theory of mind impairments in social anxiety disorder,” *Behavior Therapy*, vol. 45, no. 4, pp. 530–540, 2014.
- [479] K. U. Yazici and I. P. Yazici, “Decreased theory of mind skills, increased emotion dysregulation and insight levels in adolescents diagnosed with obsessive compulsive disorder,” *Nordic J. Psychiatry*, vol. 73, no. 7, pp. 462–469, 2019.
- [480] M. Brüne, ““theory of mind” in schizophrenia: A review of the literature,” *Schizophrenia Bulletin*, vol. 31, no. 1, pp. 21–42, 2005.
- [481] C. Montag, A. Ehrlich, K. Neuhaus, I. Dziobek, H. R. Heekeren, A. Heinz, and J. Gallinat, “Theory of mind impairments in euthymic bipolar patients,” *J. Affective Disorders*, vol. 123, no. 1–3, pp. 264–269, 2010.
- [482] A. P. Atkinson, “Impaired recognition of emotions from body movements is associated with elevated motion coherence thresholds in autism spectrum disorders,” *Neuropsychologia*, vol. 47, no. 13, pp. 3023–3029, 2009.
- [483] S. K. Jarayam, M. Masmoudi, and M. Hammami, “A comparative study of autistic children emotion recognition based on spatio-temporal and deep analysis of facial expressions features during a meltdown crisis,” *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 83–125, 2021.
- [484] R. Savery and G. Weinberg, “Robots and emotion: A survey of trends, classifications, and forms of interaction,” *Advanced Robotics*, vol. 35, no. 17, pp. 1030–1042, 2021.
- [485] F. Cavallo, F. Semeraro, L. Fiorini, G. Magyar, P. Sinčák, and P. Dario, “Emotion modelling for social robotics applications: A review,” *J. Bionic Engineering*, vol. 15, no. 2, pp. 185–203, 2018.
- [486] S. Marcos-Pablos and F. J. García-Peñalvo, “Emotional intelligence in robotics: A scoping review,” in *Proc. Int. Conf. Disruptive Technologies, Tech Ethics and Artificial Intelligence*, 2021, pp. 66–75.
- [487] A. Agnihotri, A. Chan, S. Hedao, and H. Knight, “Distinguishing robot personality from motion,” in *Companion of the 2020 ACM/IEEE Int. Conf. Human-Robot Interaction*, 2020, pp. 87–89.
- [488] J. R. Smith, D. Joshi, B. Huet, W. Hsu, and J. Cota, “Harnessing AI for augmenting creativity: Application to movie trailer creation,” in *Proc. ACM Int. Conf. Multimedia*, 2017, pp. 1799–1808.
- [489] National Safety Council, “Cost of Fatigue at Work,” <https://www.nsc.org/work-safety/safety-topics/fatigue/calculator/cost>, 2020.
- [490] A. Baghdadi, L. A. Cavuoto, A. Jones-Farmer, S. E. Rigdon, E. T. Esfahani, and F. M. Megahed, “Monitoring worker fatigue using wearable devices: A case study to detect changes in gait parameters,” *J. Quality Technology*, pp. 1–25, Aug. 2019.
- [491] M.-H. Sigari, M. Fathy, and M. Soryani, “A driver face monitoring system for fatigue and distraction detection,” *Int. J. Vehicular Technology*, vol. 2013, 2013.
- [492] S. Karvekar, M. Abdollahi, and E. Rashedi, “A data-driven model to identify fatigue level based on the motion data from a smartphone,” in *Proc. IEEE Western New York Image and Signal Processing Workshop (WNYISPW)*, 2019, pp. 1–5.
- [493] G. Sikander and S. Anwar, “Driver fatigue detection systems: A review,” *IEEE Trans. Intelligent Transportation Systems*, vol. 20, no. 6, pp. 2339–2352, 2019.
- [494] N. I. Vargas-Cuentas and A. Roman-Gonzalez, “Facial image processing for sleepiness estimation,” in *Proc. Int. Conf. Bio-engineering for Smart Technologies (BioSMART)*, 2017, pp. 1–3.
- [495] Y. Yamada and M. Kobayashi, “Fatigue detection model for older adults using eye-tracking data gathered while watching video: Evaluation against diverse fatiguing tasks,” in *Proc. IEEE Int. Conf. Healthcare Informatics (ICHI)*, 2017, pp. 275–284.
- [496] —, “Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults,” *Artificial Intelligence in Medicine*,

vol. 91, pp. 39–48, 2018.

- [497] T. D. O. Noakes, "Fatigue is a brain-derived emotion that regulates the exercise behavior to ensure the protection of whole body homeostasis," *Frontiers in Physiology*, vol. 3, p. 82, 2012.



James Z. Wang (Senior Member, IEEE) is a Distinguished Professor of the Data Science and Artificial Intelligence area and the Human-Computer Interaction area of the College of Information Sciences and Technology at The Pennsylvania State University. He received a bachelor's degree in mathematics *summa cum laude* from the University of Minnesota (1994), and an MS degree in mathematics (1997), an MS degree in computer science (1997), and a PhD in medical information sciences (2000), all from Stanford University. His research interests include

affective computing, image analysis, image modeling, image retrieval, and their applications. He was a visiting professor at the Robotics Institute at Carnegie Mellon University (2007-2008), a lead special section guest editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence (2008), and a program manager at the Office of the Director of the National Science Foundation (2011-2012).



Sicheng Zhao (Senior Member, IEEE) is currently a Research Associate Professor at Tsinghua University. He received a PhD from Harbin Institute of Technology in 2016. He was a Visiting Scholar at National University of Singapore from 2013 to 2014, a Research Fellow at Tsinghua University from 2016 to 2017, a Postdoc Research Fellow at University of California, Berkeley from 2017 to 2020, and a Postdoc Research Scientist at Columbia University from 2020 to 2022. His research interests include affective computing, multimedia, and computer vision.



Chenyan Wu is a doctoral candidate in the Informatics program of the College of Information Sciences and Technology at The Pennsylvania State University. He earned a B.E. degree in Electronic Information Engineering from the School of the Gifted Young, University of Science and Technology of China in 2018. He has worked as an intern at Amazon Lab126 (Bellevue, WA), Microsoft Research Asia (Beijing, China), and SenseTime Research (Shenzhen, China). His research interests are affective computing, computer vision, and machine

learning.



Reginald B. Adams, Jr. is a Professor of Psychology at The Pennsylvania State University. He received his PhD in social psychology from Dartmouth College in 2002. He is interested in how we extract social and emotional meaning from nonverbal cues, particularly via the face. His work addresses how multiple social messages (*e.g.*, emotion, gender, race, age, etc.) combine across multiple modalities and interact to form unified representations that guide our impressions of and responses to others. Although his questions are social psychological in origin, his research draws upon visual cognition and affective neuroscience to address social perception at the functional and neuroanatomical levels. Before joining Penn State, he was awarded a National Research Service Award (NRSA) from the National Institute of Mental Health to train as a postdoctoral fellow at Harvard and Tufts Universities. His continuing research efforts have been funded through NSF and the National Institute on Aging and the National Institute of Mental Health of the National Institutes of Health.



Michelle G. Newman is a Professor of Psychology and Psychiatry, and Director of the Center for the Treatment of Anxiety and Depression at the Pennsylvania State University. She received her PhD in clinical psychology from the University of Stony Brook in 1992 and completed a postdoctoral Fellowship at Stanford University in 1994. Dr. Newman has conducted basic and applied research on anxiety disorders and depression and has published over 200 papers on these topics. She is the past editor of Behavior Therapy and is currently Associate Editor of

Journal of Anxiety Disorders. She is also the recipient of the APA Division 12 Turner Award for distinguished contribution to clinical research, APA Division 29 Award for Distinguished Publication of Psychotherapy Research, ABCT Outstanding Service Award, APA Division 12 Toy Caldwell-Colbert Award for Distinguished Educator in Clinical Psychology, and Raymond Lombra Award for Distinction in the Social or Life Sciences. She is also a Fellow of American Psychological Association Divisions 29 and 12, the Association for Behavioral and Cognitive Therapies, and American Psychological Society.



Tal Shafir graduated from law school at the Hebrew University of Jerusalem. She then studied dance-movement therapy at the University of Haifa and completed her PhD in neurophysiology of motor control and two postdoctoral fellowships: in brain-behavior interactions in infants, and in affective neuroscience, all at the University of Michigan. She developed research on movement-emotion interaction and its underlying brain mechanisms, behavioral expressions, and therapeutic applications. Shafir, certified also in Laban Movement Analysis, was the main editor of *The Academic Journal of Creative Arts Therapies*, and *Frontiers in Psychology* research topic: 'The state of the art in creative arts therapies.' She has been serving on The American Dance Therapy Association (ADTA) research committee since 2016 and was the recipient of the ADTA 2020 Innovation Award.



Rachelle Tsachor is an Associate Professor of Movement at the University of Illinois at Chicago. She is certified in Mind-Body Medicine (CMBM), Movement Analysis (CMA), and Somatic Movement Therapy (RSMT-ISMETA). Her research investigates body movement to bring a human, experiential understanding to how movement affects our lives. Tsachor analyzes patterns in moving bodies in diverse projects, researching movement's effects on our brains, emotions, health, and learning. She is co-PI on an NSF-Funded project STAGE and the UI Presidential Initiative for the Young People's Science Theater: CPS and UIC Students Creating Performances for Social Change. Both initiatives bring Mind/Body methods into Chicago Public Schools' that educate primarily students of color to support learning in embodied ways, so students view themselves as potential contributors of creative solutions.