Review and Challenges of Technologies for Real-Time Human Behavior Monitoring

Sylmarie Dávila-Montero, *Student Member, IEEE*, Jocelyn Alisa Dana-Lê, Gary Bente, Angela T. Hall, and Andrew J. Mason, *Senior Member, IEEE*

***Abstract*—A person's behavior significantly influences their health and well-being. It also contributes to the social environment in which humans interact, with cascading impacts to the health and behaviors of others. During social interactions, our understanding and awareness of vital nonverbal messages expressing beliefs, emotions, and intentions can be obstructed by a variety of factors including greatly flawed self-awareness. For these reasons, human behavior is a very important topic to study using the most advanced technology. Moreover, technology offers a breakthrough opportunity to improve people's social awareness and self-awareness through machine-enhanced recognition and interpretation of human behaviors. This paper reviews (1) the social psychology theories that have established the framework to study human behaviors and their manifestations during social interactions and (2) the technologies that have contributed to the monitoring of human behaviors. State-of-the-art in sensors, signal features, and computational models are categorized, summarized, and evaluated from a comprehensive transdisciplinary perspective. This review focuses on assessing technologies most suitable for real-time monitoring while highlighting their challenges and opportunities in near-future applications. Although social behavior monitoring has been highly reported in psychology and engineering literature, this paper uniquely aims to serve as a disciplinary convergence bridge and a guide for engineers capable of bringing new technologies to bear against the current challenges in real-time human behavior monitoring.**

***Index Terms*—Computational models, human behaviors, multisensor modalities, automated and real-time monitoring.**

I. Introduction

HUMANS are a highly social species expressing individual behaviors that, when accumulated, create the social environment in which society operates. In general, human behavior is driven by personal factors such as thoughts and emotions that are influenced by our environment and social interactions. Social interactions, constructed by the behaviors of two or more individuals, are highly complex and play an important role in our health and survival [1]. Indeed, it has been shown that social interactions, both quantity and quality, impact our health behavior, mental health, physical health, and mortality risk [2], [3]. For example, workplace environments are of high interest for studying the influence of human behavior on social interactions, and vice versa. This is often due to the diversity in demographics and knowledge composition within a workplace, which is essential for problem-solving and innovation [4], [5]. However, diversity also presents challenges at the sharing and interaction level [6] that ultimately impact the well-being and productivity of individuals involved [7]. Moreover, several studies have indicated that low quality of social interactions is associated with a variety of health conditions, including higher risk of cardiovascular disease, compromised immunity, and increased risk of depression, among others [8]–[11]. Thus, mechanisms promoting healthier behaviors and interactions provide tremendous benefits to healthcare, e.g., to understand and treat depression and anxiety disorders, among others. A key step toward such mechanisms is to apply technology to the study and monitoring of human behaviors.

Traditionally, human behaviors have been studied using expert observations (including experiments) and/or through surveys given to the individuals involved in the interaction. However, these methods are not suitable for the real-time recognition and interpretation of human and social behaviors, which would enable feedback aiming to improve those behaviors as they occur. Real-time feedback is necessary to improve the situational awareness [12]–[15] that plays an important role in our daily lives and affects the quality of social interactions [16], [17]. Fig. 1 illustrates the concept of real-time human behavior monitoring technologies improving situational awareness during social interactions. Here, the technology captures and interprets human behavior data during interactions and instantaneously provides informative feedback to each individual regarding the social ecosystem. Feedback messages could relate to conversation patterns and dynamics (who is dominating the interaction, interruptions, etc.), strength of rapport (synchronicity between communication partners), or even individual levels of emotional arousal. However, the information that could be fed back greatly depends on the capabilities of the human behavior monitoring technologies, which are reviewed and analyzed in this paper.

Manuscript received September 3, 2020; revised January 10, 2021; accepted February 14, 2021. Date of publication February 19, 2021; date of current version March 30, 2021. This work was supported by the U.S. National Science Foundation Graduate Research Fellowship under Grant DGE-1424871. *(Corresponding author: Sylmarie Dávila-Montero.)*

Sylmarie Dávila-Montero and Andrew J. Mason are with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48824 USA (e-mail: davilasy@msu.edu; mason@msu.edu).

Jocelyn Alisa Dana-Lê is with the Department of Management, Michigan State University, East Lansing, MI 48824 USA (e-mail: jadl@msu.edu).

Gary Bente is with the Department of Communication, Michigan State University, East Lansing, MI 48824 USA (e-mail: gabente@msu.edu).

Angela T. Hall is with the School of Human Resources and Labor Relations, Michigan State University, East Lansing, MI 48824 USA (e-mail: athall@msu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TBCAS.2021.3060617.

Digital Object Identifier 10.1109/TBCAS.2021.3060617

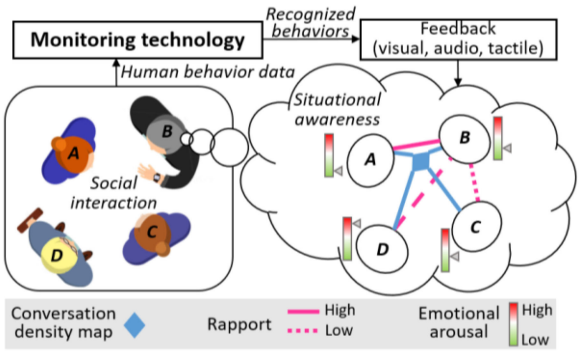

Fig. 1. Concept of enhanced social awareness using real-time monitoring of human behaviors during social interactions in an example work team meeting. Monitoring technology can instantaneously measure informative metrics of human behavior such as who is dominating the conversation (conversation density map), status of interpersonal rapport, and levels of individual emotional arousal. Recognized behaviors can then be fed back to each individual, enhancing their real-time situational awareness.

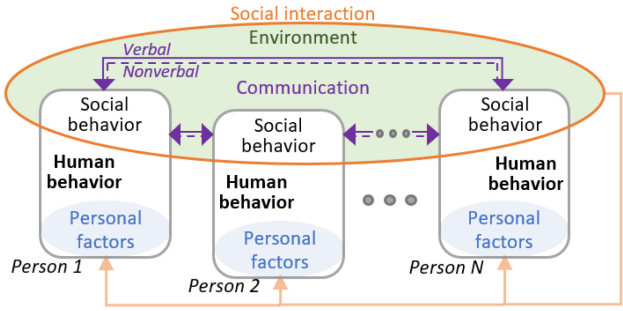

Fig. 2. Diagram describing the effectors of human behavior and their dynamics. In short, given an environment, personal factors influence human behaviors, which influence our social behaviors affecting how we communicate during social interactions. In a reciprocal loop, the elements involved in social interactions influence back our personal factors, which influence our human behaviors and so on.

To provide a clear understanding of state-of-the-art technologies for human behavior monitoring and promote convergence research into new technologies that can overcome current challenges, this paper provides an extensive review of the literature associated with monitoring human behavior. While multiple reviews of human behavior monitoring have been published [18]–[27], this paper uniquely presents a comprehensive, transdisciplinary, perspective with a focus on identifying critical design considerations in real-time human behavior monitoring systems. Starting with an overview of social psychology theories that have established the framework to study human behaviors and their manifestations during social interactions, this paper then establishes a taxonomy of human behavior monitoring technologies based on these psychological theories. Neoteric elements of this review are an insightful categorization of sensors and an informative analysis of signal characteristics, features, and computational models that have been reported. Moreover, this review focuses on sensor hardware and real-time signal processing technologies that have proven most effective for embedded monitoring of human behaviors while highlighting challenges and opportunities in near-future wearable applications. The design of real-time human behavior monitoring technologies requires a multidisciplinary effort that includes electrical engineering, data science, cognitive engineering, psychology, neuroscience, and communication science. To enable convergence across these disciplines, this paper brings together the important concepts that empower embedded sensor systems engineers to integrate human behavior analysis into next generation wearable health monitoring systems.

This review paper is organized as follows: Section II presents the psychological theories and concepts that underpin the analysis of human behaviors; Section III presents the taxonomy used for the categorization of the reviewed literature; Section IV presents a summary of sensor systems and a categorization of the different sensor types used in the monitoring of human behaviors; Section V presents sensor signal characteristics and features that are informative of human behaviors; Section VI presents a summary of the computational methods used for the classification of human behaviors; and Section VII assesses the current challenges and opportunities in human behavior monitoring, followed by the conclusion in Section VIII.

II. Theories and Concepts of Human Behaviors

A. Human Behavior and Its Effectors

Behavior is generally defined as the "observable consequences of the choices a living entity makes in response to external or internal stimuli" [28]. Internal stimuli could be a person's thoughts, memories, perceptions, or attitudes, while external stimuli come from the environment the person interacts with, including social interactions. In humans, depending on the level of situational and personal awareness that they possess, responses to external and internal stimuli (effectors of human behavior) can be voluntary or involuntary. Fig. 2 shows the dynamics of the effectors of human behavior, which can include personal factors and components of social interactions.

As illustrated in Fig. 2, personal factors are inside the person. They can come from a person's biology or psychology. Personal factors that come from a person's psychology are in the mind and are not externally observable; however, the behaviors a person expresses because of the influences of their psychological personal factors are directly observable. Social behaviors, a subset of human behaviors that are specifically directed at other people or that involve a social action, are directly observable. Communication, both verbal and nonverbal, is a vital aspect of social interactions and also directly observable. As illustrated in Fig. 2, social behaviors strongly influence communication dynamics during a social interaction while, in a reciprocal loop, our social behaviors get influenced by our social interactions. Part of this idea is captured by the well-established *Social Cognitive Theory (SCT)*, which contends that individuals' perceptions of their environment can influence their emotional, physiological, and behavioral reactions [29], [30], subsequently influencing future behaviors in a reciprocal loop.

To properly understand the technology developed to monitor human behavior, one must first understand the personal factors that underpin human behaviors and the theories and concepts that have guided the psychological study of social interactions.

These two topics are briefly summarized below to provide a scholarly foundation for our subsequent review of human behavior monitoring technologies.

B. Personal Factors

The psychological factors that have been commonly studied that contribute to human behavior are affect and dispositions. Affect generally means anything related to a person's emotions or moods, and it can be divided into two categories: states and traits. State affect is an emotion or mood that is experienced in a certain moment of time, whereas trait affect is a more enduring part of one's personality. Emotions are mental and physiological experiences of feeling that are acutely experienced (intense) and discrete in that they have a beginning and an end point, while moods refer to the positive or negative feelings that are in the background of our everyday experiences; these are diffuse (not acutely experienced) and longer-lasting states than emotions; however, they are not as enduring as trait affect. Trait affect is part of one's personality – it is a tendency to experience certain emotions and moods in general. For example, someone might have trait negative affectivity, which is the tendency to experience negative moods and emotions more often than others. Together, these states of emotional experiences and traits constitute affect.

A disposition in the social sciences is thought of as a natural proclivity (biological or psychological) to respond to situations in a particular way. Because dispositions are "natural" and inherent in the person, they are thought to be the most stable and enduring phenomenon studied in the psychological sciences that are discussed in this review (i.e., more enduring across time than state affect, attitudes, and behaviors). However, despite their stability across time, dispositions do not relate to behavior with perfect consistency because there are environmental factors that also influence behavior. For example, a person might have a biological disposition to develop a psychological disorder, but through certain training environments like therapy, they are able to override their disposition to develop the disorder. As another example, someone may be genetically predisposed to have a reserved personality, but they are put in social environments that constantly require them to talk to others, so they override their genetic predisposition. Dispositions influence human behavior more when the situation is weak, like when a person is in a casual social interaction. On the other hand, dispositions influence human behavior less when the social situation is strong and enforces certain norms, such as a professional environment in which all individuals are expected to behave in a certain way regardless of their personalities [31]. In this review, we focus on personality traits, which are influenced by dispositions as well as by the environment [32].

Lastly, another important personal factor that impacts behavior is attitudes. An attitude is a psychological tendency to evaluate a particular target with some degree of favor or disfavor [33]. The "target" could be another person or a non-living thing such as a food, brand, or idea. An attitude, at its core, is an evaluation. Thus, it differs from affect and dispositions. Whereas affect could include an emotion that arises in response to a target, an attitude is a feeling towards the target and a set of judgments about the target. Attitudes are more enduring than a state but less enduring than a trait or disposition. Fig. 3 shows the relationship between these personal factors and time. This relationship is typical across much of the psychology and organizational behavior literature. The duration of these factors and the interactions between them play an important role in understanding how technology can be used to understand human behaviors.

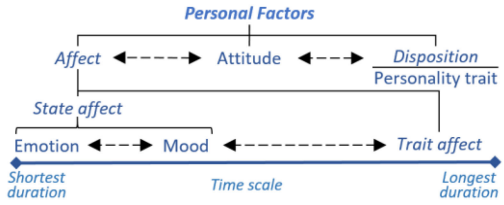

Fig. 3. Diagram describing the personal factors that influence human behavior and how they manifest through time. Personal factors include affect, attitudes, and dispositions, of which, affect and dispositions are the most studied. Affect is divided into states and traits. State affect is related to acute emotions and mood, in contrast to trait affect which is related to a human's disposition to experience positive or negative emotions and is a more enduring part of human personality.

We contend that there is a lack of research using behavior monitoring technologies to study the role of attitudes in human behavior during social interactions. Thus, next, we focus on reviewing psychological theories that delve specifically into the explanation of state affect and personality traits.

1) State Affect-Related Theories: Discrete, acute emotions often provoke a person to mentally narrow in on a specific action or set of actions. For example, the experience of fear leads to the activation of thoughts in the mind about defending oneself or running away (also known as "fight-or-flight" response), and the experience of interest can activate a person's thoughts aimed at exploring and taking in new information [34]. "Activations of thought" driven by emotions can occur subconsciously. Indeed, the body mobilizes physiological resources to complete these actions without the person's conscious awareness.

Based on the explored idea that emotions reflect responses of the sympathetic nervous system [35], the *Polyvagal Theory* explains how state affect alters brain processes and biological processes that occur in the rest of the body [36], [37]. In addition, this theory provides insights about the relationship between measurable physiological states, linked to the autonomic and central nervous systems, and the resulting human behavior, suggesting a bidirectional relationship between the brain and the body. It also suggests that the environment affects behaviors that consequently alter physiological states. Thus, monitoring changes in the physiological states of the human body, such as respiration rate, heart rate, and perspiration rate, among others, can provide insights into the affective state of an individual [26]. Likewise, monitoring environmental conditions can provide information on how the environment influences emotional states and other factors.

Emotions can be-understood to fall somewhere along two orthogonal dimensions: (1) of how pleasurable the emotion is, and (2) of how much arousal or activation the emotion involves.

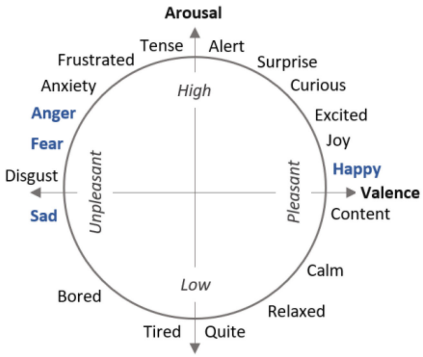

Fig. 4. A typical circumplex model of affect that describes affective states using two fundamental neurophysiological systems: valence and arousal. Valence describes the level of pleasure or displeasure of an emotion, while arousal describes its level of activation. Emotions in blue color represent the four most commonly studied emotions in affective computing.

As shown in Fig. 4, emotions are commonly arranged in a *circumplex model of affect* [38], according to where they fall on both dimensions. For example, excitement is an emotion that is pleasurable and high on arousal, whereas calmness is an emotion that is pleasurable and low on arousal. Anger and fear are unpleasant, high on arousal emotions close together on the circumplex, whereas boredom is a low-arousal unpleasant emotion. The circumplex model of affect is a mainstream and well-established theory. However, other dimensional models of emotions have also been used to study emotional states, such as the *Pleasant, Arousal, and Dominance (PAD) emotional state model* [39] that, in addition to modeling emotions in a valence-arousal scale, contains a dominance dimension representing the controlling nature of an emotion. The *Plutchik's model* [40] is another dimensional model of emotions that organizes discrete emotions from the most basic to the most complex ones. In the field of affective computing, happiness, sadness, anger, and fear are the four most commonly studied emotions.

A discrete emotion can affect someone's response to a social interaction whether the emotion was caused by that interaction or not. The well-established framework of *Emotions As Social Information (EASI) model* [41], [42] asserts that emotions serve a social function by relaying information when they are expressed. For example, if a person is late to a meeting with a coworker and the coworker appears to be angry, this provides information that leads to certain inferences, such as the inference that the person was late, the inference that the behavior of being tardy was inappropriate, and the inference that the person should strive to arrive earlier in the future [41]. The information that emotions relay to others in social interactions is valuable for adjusting future behavior.

2) Personality Trait-Related Theories: The dominant theory in the organizational sciences used to taxonomize personality is the Five-Factor Model (FFM) of personality, also known as the "Big Five." The "Big Five" factors of personality can be abbreviated with the acronym "OCEAN": Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Openness involves being open to new experiences, unconventional, nonconforming, creative, and imaginative, while conscientiousness is the "tendency toward being dependable, disciplined, purposeful, organized, and achievement-oriented" [43]. Extraversion, in the sense of the FFM, is the "tendency to be social, talkative, energetic, and active" [43]. It was found that among the personality factors, extraversion has the strongest relationship with leadership (both being recognized as a leader by others and being effective as a leader) [44]. On the other hand, agreeableness tends to be a catchall factor related to aspects of personality that are likable and harmonious with others, such as being trusting of others, polite, empathetic, and compliant [45]. The last one of the "Big Five" is Neuroticism, which is often labeled as its opposite instead, emotional stability. Those who are high on neuroticism are more likely to experience negative emotions like anxiety, anger, irritation, frustration, and jealousy. On the other hand, having low neuroticism or high emotional stability means that a person tends to be more even-keeled, calm, and unwavering (not necessarily positive or enthusiastic).

There are many other taxonomies of personality, such as the HEXACO model [46] which breaks the agreeableness factor of the FFM into agreeableness and humility. The Dark Triad is another taxonomy that has only three undesirable personality traits: narcissism, Machiavellianism, and psychopathy [47]. Another trait is *locus of control*, which describes the extent to which individuals believe that they control their own outcomes as opposed to having their successes and failures determined by external forces [48]. So far, the traits covered by the FFM and the *locus of control* have been studied using human behavior monitoring technologies. In general, the activation of these traits during social interactions contribute to observable social behaviors that make up the social environment that people operate in.

C. Communication

One of the most important factors influencing our human behavior is the social behavior of our interaction partners. We might think of it as a situational factor, but this would ignore the dynamic nature of mutual adaption within the communication process. By definition, "communication is a transactional process in which people generate meaning through the exchange of verbal and nonverbal messages in specific contexts, influenced by individual and societal forces and embedded in culture" [49]. Verbal communication refers to the use of spoken and written language (words). It is usually organized in distinct on-off patterns of messages or utterances with iterating sender and receiver (speaker, listener) roles. The use of words requires a shared explicit code usually to be found in a dictionary. Spoken language, however, can also carry implicit, so-called paraverbal, information, as for instance encoded in the floor possession and pausing or in prosodic features such as pitch, speed, and volume of the vocal output.

Nonverbal communication comprises all aspects of communication that are not encoded into words. In stark contrast to verbal communication, nonverbal communication is continuous, i.e. always on, and largely implicit, i.e. it lacks a dictionary

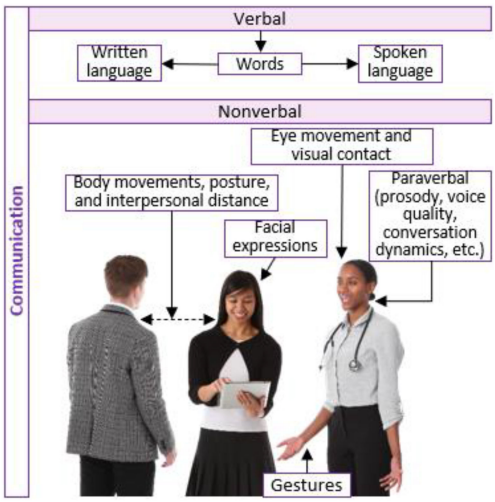

Fig. 5. Elements of communication relevant to social behaviors during interactions modeled by an exchange of verbal and nonverbal messages. Verbal communication involves the use of words through written or spoken language. Nonverbal communication involves the use of gestures, facial expressions, paraverbal communication, eye movements, visual contact, body movements, posture, and interpersonal distance.

and is produced and processed widely automatically and unconsciously. Therefore, it is hard to control and its effects impose on the observer with an irrefutable force. Even a lack of nonverbal expressions is interpreted by the observer, for instance, as disinterest. In this sense "we cannot not communicate" [50]. It has been argued that our social perception and impression formation is much more dependent on nonverbal messages than on verbal behavior and that nonverbal communication can be conceived as meta-communicative [50], in the sense that it even largely defines how we understand and interpret the spoken words. Thus, even in the presence of verbal communication, successful communication largely depends on the efficient use of nonverbal communication channels [51].

As nonverbal communication largely withdraws deliberate manipulation, it is supposed to provide information about unobservable processes such as the individual's emotional state, intentions, personality traits, etc. [52]. The view that nonverbal communication is a reliable source of true information, although still under debate [53], has made it the focus of study of many areas dedicated to understand social behavior and the human mind. For example, the social signal processing [54], [55] and affective computing [56] literature, areas of engineering and computer science that study social interactions and human emotions, respectively, focus on the messages produced by nonverbal communication channels. They are better known in those areas as "social signals", a notion that was first introduced in the field of computational social science and organization engineering [57]. Thus, we focus on reviewing the most commonly used nonverbal communication channels.

Due to its unique level of complexity, the analysis of nonverbal communication poses considerable methodological challenges [58]. Nonverbal communication implies multiple channels and serves various functions. As illustrated in Fig. 5, nonverbal communication includes gestures, body movements and postures [59]–[62], facial expressions [63], [64], and eye gaze [61], among others. As most authors, we here subsume paraverbal communication, such as prosody, pitch, volume, and intonation [65], [66], under the broader construct of nonverbal communication.

Nonverbal communication comprises attentional functions, interpretations and most importantly the regulation of interpersonal relations. We distinguish three, distinct, yet interdependent functions of nonverbal communication [67]: (1) discourse functions, (2) dialog functions and (3) socio-emotional functions that influence our social behaviors.

Discourse functions are closely related to speech production and understanding. Emblems, pointing gestures, illustrative gestures and beat gestures belong to this functional category [68], as well as prosodic aspects, such as pausing and variations in voice pitch and volume. In general, discourse functions influence aspects of interpersonal communication and engagement that includes listener attention, interest, understanding, and interpretation.

Dialogue functions include turn-taking signals (e.g. eye contact, raise of voice, pausing) and back-channel signals (e.g. head nods, 'uh-huh', etc.), which serve to smooth the flow of interaction when exchanging speaker and listener roles [69]. In addition, dialogue functions influence aspects of social interactions that include communication patterns, conversation dynamics, and level of interaction between individuals.

Socio-emotional functions of nonverbal behavior include the communication of emotions and interpersonal attitudes and their regulation, which are crucial for establishing rapport. Whether we harmonize in an interaction, take others' perspectives or are capable of establishing a smooth flow of interaction very much depends on the exchange of those socio-emotional cues. Socio-emotional functions are not independent from dialogue and discourse functions of nonverbal behavior. A smooth flow of the conversation will most likely influence the interaction climate in a positive way. Power relations are evident in body postures, eye contact, voice amplitudes and more [70]. Harmony or interpersonal rapport shows in expressiveness or responsiveness [71] as well as in mutual attentiveness (body orientation, eye contact), reciprocal positivity (smiles, interpersonal distance, body lean, and orientation) and in behavioral coordination (motor mimicry, posture sharing and synchrony, and activity entrainment). Thus, monitoring nonverbal messages provides insights into the human and social behaviors being displayed given an environment [72], [73].

III. Monitored Human Behaviors

The literature surrounding technology for human behavior monitoring is vast and varied. Even focusing down to only technologies with the potential to advance automatic and/or real-time monitoring, as we have chosen for this paper, synthesizing reported technologies to enable an analytical perspective motivates the creation of a classification system. Consider that an underlying goal of this paper, and indeed of most of the efforts reviewed herein, is to enhance the potential for technologies

TABLE I
TAXONOMY SUMMARIZING HUMAN BEHAVIORS MONITORED USING SENSOR TECHNOLOGIES

Effector classes (complexes)	*Behavioral elements (aspects/components/dimensions)*	*References*
Emotions	Dimensional: valence, arousal, potency Categorical (basic emotions): happy, angry, sad, quiet, disgust, anxiety, surprise Others: curiosity, boredom, uncertainty, puzzlement	[13], [23], [110]–[115], [154], [156]–[158], [91], [159]–[162], [166]–[169], [177], [179], [92], [180]–[189], [93], [190]–[199], [94], [200], [95]–[97], [101]
Personality factors	Personality traits: leadership emergence, openness, conscientiousness, extraversion, agreeableness, and neuroticism Person Perception Dimensions: valence, dominance, activity Others: empathy, honesty	[79], [86], [201]–[209], [116], [119]–[123], [163], [168]
Social interactions	Cooperation or collaboration, agreement and disagreements, attraction, interest, attention, emphasis, vigilance, group performance, cohesion, communication patterns and dynamics, level of interaction, rapport	[13], [14], [87]–[89], [124]–[130], [76], [134]–[137], [155], [164], [168], [210]–[212], [79], [213]–[222], [81], [223], [224], [82]–[86]

that augment human capability, toward a future of increasingly effective human-machine interactions. To further promote this human-centered approach, we established a taxonomy for behavior sensing technology that is based on the relevant psychological theory summarized in Section II. Specifically, our taxonomy assigns technologies to the human behavior effectors that they target, and it defines three effector classes that encompass the reviewed literature, as shown in Table I. The defined effector classes cover personal factors as well as social interaction factors observed through nonverbal communication channels, all of which influence human behavior.

In brief, Table I assigns the emotions effector class to works that concentrated on recognizing categorical and dimensional emotional structures, most of which focus on understanding an individual's emotional state, rather than the dynamics of emotional expression and exchange during a social interaction. The personality factors class was allocated to works related to personality traits and person perception dimensions as well as to works centered on the detection of empathy and honesty. Finally, the social interactions class was assigned to works covering aspects of interpersonal communication and engagement such as levels of interest, level of cohesion, and communication dynamics.

We will maintain the taxonomy established by Table I throughout this review. The taxonomy will be used to discern similarities and differences in the various sensors, signal features, and computational models employed for monitoring within these prescribed human behavior effector classes.

IV. SENSORS FOR MONITORING HUMAN BEHAVIORS

The activation of emotions, personality factors, and aspects of social interactions manifests in the form of physiological processes and nonverbal messages that can be captured using sensor technologies, including transducers (recognition elements) and their essential interface circuits. Toward the goal of real-time data collection and analysis of human behaviors, a wide range of wearable monitoring systems have been developed and a variety of sensor modalities employed.

A. Wearable Sensor Platforms

In the 1990s, the early development of wearables to study human behaviors were focused on aspects of social interactions. These initial systems, still used nowadays, employed InfraRed (IR) and/or quasipassive radio frequency (RF) sensor modules to track position and proximity among individuals wearing these devices [74]–[76]. In an effort to create wearable systems with the ability of capturing more informative data about human behaviors, research groups started working on the integration of multiple sensing modalities. One of the earliest initiatives was the MIThril project pioneered by A. Pentland [77]. The MIThril project focused on developing a "practical, modular system of hardware and software for research in wearable sensing and context-aware interaction" [77]. With the introduction in the early 2000s of the personal digital assistant (PDA) devices, the MIThril project first developed a modular wearable system comprised of a variety of sensors such as accelerometers, InfraRed (IR) active tag readers, GPS units, analog microphones, 2-channel electromyography (EMG) sensors, 2-channel electrodermal activity (EDA) sensors, and skin temperature monitors [78]. The sensors were wired to a PDA intended to perform real-time processing and communicate with other units of the same kind through Wi-Fi. However, there seem to be no reports of data collected using this system. In an effort to study communication patterns of groups of people during meetings in real time, Eagle and Pentland [79], designed a wearable system employing a headset microphone connected to individuals' PDAs to allow streaming of high quality audio signals over a network, also with the choice of storing the audio locally on the device for post-processing. Conversations were detected in all streamed audio signals and conversation features extracted, including inferring the proximity among participants. Later, the same research group made use of the UbER-Badge [80], a device with a microphone, a two-axis accelerometer, and a forward-oriented IR transceiver, to measure human interest levels during interactions in a conference meeting, all among dyads [81], [82]. With a similar system called the Sociometer, people involved in an interaction were identified and through audio signals, conversation dynamics studied [83]. This version of the Sociometer was later optimized. Modified versions of the Sociometer have been known as the Sociometric badge [84]–[86], Open badge [87], and Rhythm badge [88], all these sensor platforms have been used in the study of social interactions.

Mobile phones have also been used as a platform for monitoring social interactions. In [89], the Bluetooth and microphone units from a mobile phone were used to detect proximity and conversation dynamics in real time to infer levels of interest in a social interaction. Moreover, they have also been used to connect with badges to display feedback information useful to improve social interactions [86], [88]. A review of additional wearable

sensors used for social interaction recognition can be found in [90].

Besides social interactions, wearables have also been used for real-time emotion recognition. In [91] and [92], accelerometer, gyroscope, ambient light, temperature and humidity sensors were integrated into a watch-like device to monitor levels of anxiety in human subjects. In an improved version that includes a MEMS microphone and a skin temperature sensor, Jiang *et al.* [93] used this wearable systems for health monitoring to study the relationship between mental health and physical health. In [94], Breeze, a wearable pendant placed around the neck, with an inertial measurement unit (IMU), was employed to measure breathing patterns as these are closely linked to emotions. The goal of the researchers was to improve the emotional states of the Breeze users by providing real-time feedback on the user's breathing patterns. Also related to the regulation of emotional states, in [95], a wearable system in the form of a glove containing an EDA, a blood volume pulse (BVP), and a skin temperature sensor was designed to continuously monitor changes in the physiological signals that could relate to emotional mental states. On the other hand, Girardi *et al.* [96] used commercially available wearable sensors to capture electroencephalography (EEG), EDA, and EMG signals to detect emotions in the arousal-valence dimensions. Also using commercially available sensors, McGinnis *et al.* [97] employed accelerometers and gyroscopes to diagnose anxiety and depression in young children. A comprehensive list of commercially available wearable physiological sensors, used especially for the monitoring of emotions, can be found in [98].

B. Categorizing Behavior Monitoring Sensors

Machine monitoring of human behavior starts with the appropriate selection of sensors. Commonly used sensors in monitoring human behavior can be grouped as sensors that capture video and images, audio, physiological, movement, orientation, proximity, and environmental signals. The selection of a sensor or multiple sensors is driven by the type of behavior to be monitored and its associated nonverbal messages and physiological reactions. Based on analysis of reviewed literature, 21 different sensors were found to have been used to monitor human behaviors. Table II lists the sensors used in the reviewed literature and the nonverbal messages, physiological reactions, and/or environmental conditions that can be captured by them. In addition, Table II summarizes information related to sensor placement and level of superficial invasiveness to the user. In our definition of superficial invasiveness, we refer to the degree at which the sensor requires to enter in contact with the body and not whether the sensor needs to be implantable in the body. In fact, to the best of our knowledge, no implantable sensor has been reported to be used in the application of human behavior monitoring. Thus, we classify the level of superficial invasiveness to the user in three categories: skin contact (sensor requires direct contact with the skin), body contact (sensor has to be placed on the body but do not requires direct contact with the skin), and no contact, with skin contact being the most invasive and no contact completely non-invasive. For sensors that require skin contact, we have indicated if they require a single point of contact or multiple points of contact. This information is useful when assessing the level of obtrusiveness of a given system or evaluating sensors for the design of wearable systems.

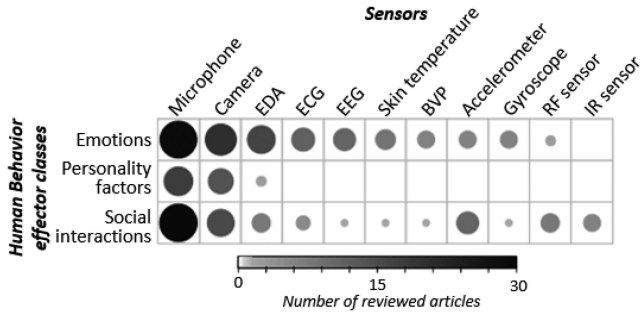

Fig. 6. Graphic representation of where the work related to human behaviors has been concentrated relative to the 11 most used sensor modalities. The most active area is monitoring of emotions, followed by the monitoring of social interactions. Microphones, cameras, and EDA sensors are the only sensing modalities used in the monitoring of all of the three effector classes, with microphones being the most common sensor modalities.

From the sensors listed in Table II, we looked at the top 11 most frequently used sensors in the literature and plotted their frequency of use with respect to the effector classes presented in Table I. The relationship between the top 11 most frequently used sensors and the effector classes is summarized in Fig. 6. In the mentioned figure, it can be observed that the monitoring of emotions has been one of the areas of most interest followed by the monitoring of social interactions, with microphone as one of the most common sensor modalities used for their study. It can be noticed that microphones, cameras, and EDA sensors are the only sensing modalities used in the monitoring of all of the three effector classes. In the cases of microphones and cameras, it is presumably because of the quality and quantity of the information that they provide, the numerous advances in the areas of speech and image processing, and advantages in terms of superficial invasiveness and placement. However, the use of cameras (to capture image and/or video) require a large data bandwidth of communication compared to microphone data. In fact, most of the literature reporting the use of video cameras for the monitoring of human behavior has been for offline applications. It is important to mention that when video and image data is included to analyze human behavior, the computational load and power consumption of the system increases [99]. In addition, it has been a topic of debate that the use of cameras to monitor human behavior presents a concern for user privacy. Thus, as video and image modalities present limitations for real-time and wearable applications, we exclude them from further analysis in this review. However, information on the use of video and image sensor modalities for the monitoring of human behaviors, including the use of facial expressions for the recognition of the emotion effector, can be found in [24], [100]. On the other hand, although the privacy issue could be argued to also apply to microphone data, in the case of monitoring human behavior as presented in this review, speech recognition is not the goal. In this area, microphones are used mostly to perform speech detection to extract acoustic features in speech (e.g., volume,

TABLE II
CATEGORIZATION OF SENSOR TECHNOLOGIES USED IN THE LITERATURE* TO MONITOR HUMAN BEHAVIOR, TOGETHER WITH ITS INFORMANTS, ASSOCIATED EFFECTOR CLASSES AND SENSOR MODALITY, AND THE LEVEL OF INVASIVENESS OF THE SENSORS RELATIVE TO THEIR PLACEMENT. ABBREVIATIONS: EMOTIONS (E), PERSONAL FACTORS, (PF), SOCIAL INTERACTIONS (SI), UNIMODAL (Uni), MULTIMODAL (MULTI)

Type	*Sensor*	*Nonverbal, physiological, & other informants*	*Effector classes*			*Sensor modality*		*Level of superficial invasiveness*	*Placement*
			E	*PF*	*SI*	*Uni*	*Multi*		
Audio	Microphone	Prosody, pitch, speech volume, intonation, turn-taking, pauses, speech duration	✓	✓	✓	✓	✓	Body contact or no contact	Chest or in front of an individual (on a table)
Video and Image	Camera	Gestures, body movements, body lean and orientation, postures, facial expressions, eye gaze	✓	✓	✓	✓	✓	No contact	In front of the individual or room view
Movement, orientation, and proximity	Accelerometer	Body movements, body lean and orientation, postures, gestures, breathing patterns	✓		✓		✓	Body contact	Chest, left wrist, belt, necklace, in the right trouser pocket, shirt pocket, or bag
	Gyroscope		✓				✓		
	Magnetometer		✓				✓		
	InfraRed (IR) sensor	Orientation (face-to-face time), proximity			✓		✓		Chest, head
	Ultrasonic sensor				✓		✓		Chest
	GPS	Proximity			✓		✓		Chest, belt, pocket, or bag
	Radio Frequency (RF) – Bluetooth included	Proximity, gestures, body movements	✓		✓	✓	✓	Body contact or no contact	Chest, belt, pocket, bag, or room
	Eye tracker (optical)	Eye gaze	✓		✓		✓		Face or in front of an individual (on a monitor)
Physiological	Blood volume pulse (BVP) /Photoplethysmography (PPG) sensor	Blood volume in arteries and capillaries, heart rate	✓		✓		✓	Skin contact – single point of contact	Wherever there is an easy access to a pulse. Fingers or earlobes are commonly used
	Respiration (RSP) sensor	Respiration rate	✓		✓		✓		Chest
	Skin temperature monitor	Skin temperature	✓		✓		✓		Any site on the body with preference in the axilla and forehead
	Electrodermal activity (EDA) /Galvanic Skin Response (GSR) sensor	Skin conductivity	✓	✓	✓	✓	✓	Skin contact – two points of contact	Fingers, palm of the hands, soles of the feet, or wrist
	Electrocardiogram (ECG)	Heart rate	✓		✓	✓	✓	Skin contact – multiple points of contact	Chest or limbs
	Electroencephalography (EEG)	Brain activity	✓		✓	✓	✓		Along the scalp
	Electroglottography (EGG)	Pitch, turn-taking, pauses, speech duration, utterances	✓				✓		Surface of the neck
	Electromyography (EMG)	Facial expressions	✓				✓		Facial muscles
	Electrooculography (EOG)	Eye gaze	✓		✓		✓		Face, around the eyes
Environment	Ambient temperature sensor	Environmental factors	✓				✓	Body contact or no contact	Wrist or in a room
	Humidity sensor		✓				✓		
	Ambient light sensor		✓				✓		

* Information presented in this table was obtained by analyzing collected information from the articles referenced in Table I.

signal energy, pitch), which can be performed in a local device before any data transmission and at reasonable computational and power consumption rates [93], [101].

Besides cameras and microphones, and in addition to EDA sensors, four physiological sensors that have been commonly used are ECG, EEG, skin temperature, and BVP sensors. The wearability of physiological sensors, which has been possible due to advancements on CMOS and circuits technologies [102], [103], has allowed the study of human behavior effectors in different scenarios. ECG, EEG, skin temperature, and BVP sensors have helped understand acute and long-term changes in the physiology of the human body that are often altered by internal and external stimuli, but that our conscious mind cannot control. All of those four physiological sensors have been used to monitor emotions and aspects of social interactions, as noted in Fig. 6. On the other hand, accelerometers, gyroscopes, IR sensors, and RF sensors are of the most frequently used sensors from the movement, orientation and proximity sensors type. While accelerometers, gyroscopes, and RF sensors have been used in the recognition of emotions, IR sensors have just been used in the monitoring of social interactions to measure proximity between individuals.

In addition to sensor modalities that are directly related to measurements of an individual, the contextual or environmental information in which signals of an individual are collected could help improve machine understanding of a behavior. Although the use of environmental sensors in the human behavior monitoring literature is scarce, these sensors are to be used to add to the

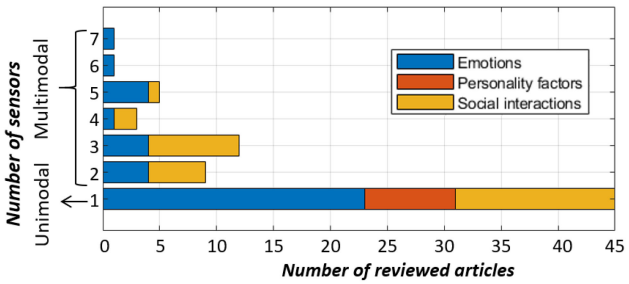

Fig. 7. Distribution of the use of unimodal and multimodal (excluding video and images) sensor modalities to monitor human behaviors. Of ∼74 reviewed works, around 59% of them rely on unimodal sensing, including all works targeting personality factors, and roughly 40% use two or more sensor modes.

contextual understanding of a behavior. For example, in [91] and [92], the use of environmental sensors such as temperature, humidity, and ambient light were used in a wearable sensing device to help determine moments of personal anxiety.

C. Analyzing Unimodal Versus Multimodal Sensor Systems

While Table II and Fig. 6 illuminate the breadth of sensors employed for behavior monitoring and their relative popularity in the literature, it is also important to consider the number of different sensor modes employed among these studies. To provide some insight on this, Fig. 7 plots the distribution of sensor modalities in relation to the identified effector classes across the articles that were analyzed. This plot shows that, of the ∼74 reviewed works, around 59% of them rely on unimodal sensing, including all works targeting personality factors. Moreover, these unimodal efforts utilize only five of the sensor types defined in Table II, namely microphones, EDA, EEG, ECG, and RF sensors. In contrast, the roughly 40% of works that were found to use two or more sensor modes, defined as multimodal in Fig. 7, utilize all sensor types listed in Table II (except for cameras, excluded from this analysis). One might expect that, as sensor technologies advance, a trend toward multimodal sensing would be evident, and our analysis supports this, showing that 66% of the multimodal works have been published since 2017, compared to only 14% of unimodal works. Multimodal sensing also makes practical sense considering that, as social individuals, humans often communicate using multimodal signals in a complementary and redundant manner. Thus, our own actions would suggest that multimodal sensor systems would be ideal for the recognition of human behaviors.

In the area of human-computer interaction, specifically in the detection of emotions, it has been recognized that multimodal systems improve the recognition rate of human behaviors when compared to unimodal approaches [56], [104], [105]. Fig. 8 presents the range, where the central red mark indicates the median accuracy, of the reported computational classification performance accuracies for both unimodal and multimodal sensor systems in the reviewed literature. Note that Fig. 8 collects information only from works that performed a classification task and reported their results using a percentage of performance accuracy. Further details on this matter and other works that reported their performance value using other metrics (e.g., root

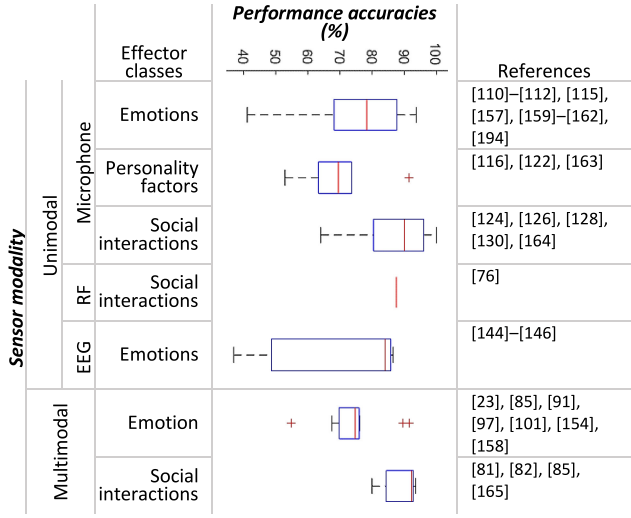

Fig. 8. Summary analysis of reported performance accuracies of unimodal and multimodal (excluding video and images) sensor systems of reviewed literature. The central mark indicates the median accuracy, and the left and right edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme accuracy values not considered outliers, while accuracy values considered outliers are plotted individually using the '+' symbol.

mean square error (RMSE), precision, recall) will be discussed in Section VI.

From Fig. 7 and Fig. 8, it can be observed that the effector classes that most utilize multimodal sensing are emotions and social interactions, a fact also noted in other behavior monitoring review papers [20], [73], [106], [107]. On the other hand, the works reviewed herein show a lack of multi-sensor modalities for monitoring personality factors. Although unimodal approaches have helped the scientific community in evaluating how information from a specific sensor contributes to understanding a certain behavior, studying the integration of multi-sensor modalities advances the development of more accurate and robust social sensing systems. Compared to unimodal sensing, multimodal sensing is still in its infancy and encounters new layers of complexity in defining and assessing accuracy. This may explain the lack of multimodal accuracy improvements observed from Fig. 8. However, multimodal systems do demonstrate less variability in accuracy, which could indicate advantages in precision and system robustness.

V. Signal Features Informative of Human Behaviors

The sensor modalities discussed in the previous sections are just one of the components necessary to capture the physiological processes and nonverbal messages associated with human behaviors. The processing of sensor signals also plays a critical role in the design of accurate real-time human behavior monitoring systems. The goal of sensor signal processing is to compute statistically identifiable signal characteristics or measurable signal properties, typically referred to as signal "features", that are informative of human behaviors.

To analyze the sensor signal processing reported in the reviewed behavior monitoring literature, works were first grouped based on their use of unimodal sensor signals and multimodal

sensor signals. Then, the unimodal works were organized by their sensor modalities and the four most used modes (excluding cameras, for reasons stated earlier), based on data in Fig. 6, were selected for further analysis. Within each modality, sensor signal processing elements such as signal characteristics, pre-processing approaches, and features were studied and summarized to illuminate the design space employed in the literature. For the analysis of signal features, reported works were grouped by their behavior effector class defined by the taxonomy established in Table I. Then, we looked for cases where the contribution of features to the recognition of a particular behavior was reported using correlation analysis or feature selection algorithms. Feature selection has two advantages: it reduces computational costs, and it removes noisy data that otherwise could degrade system performance.

The understanding gained from the analysis of unimodal sensor signal processing elements was then applied to make a qualitative assessment of their utility and design considerations in multimodal systems. Finally, we attempted to integrate this information with an analysis of the limited works presenting signal processing for multimodal systems. This effort allowed us to make the summary observations presented at the end of this section and in Section VII that may be helpful for the design of future real-time human behavior monitoring systems.

A. Audio Signals

Audio signals collected from microphones are sound waves converted into electrical energy that, when employed in human behavior recognition systems, are typically used to monitor paraverbal communication. Audio signals used to monitor paraverbal communication are usually collected using a minimum sampling rate of 8 kHz, but rates up to 44.1 kHz have also been reported. The use of higher sampling frequencies provides better signal resolution, but it is not necessary for the extraction of the acoustic features of interest. The processing of audio signals is mainly composed of four parts: speech detection, speech segmentation, signal pre-processing, and feature extraction. Thus, identifying levels of noise, periods of silence, and periods of speech becomes a key task to ultimately extract accurate features and associate them to behaviors of interest. In real-time processing, audio signals are processed in frames of $\sim$30ms to $\sim$80ms, often with overlaps between each consecutive frame. These frames of data are used to detect speech. In general, after detecting speech in the audio signal, audio segmentation is performed. Audio segmentation refers to the task of dividing the audio signal into acoustic segments from which acoustic features will be extracted [108].

Typically, in the area of human behavior monitoring, audio segmentation has been done in two ways, through an utterance-based approach or a windowing-based approach. Utterance-based approach includes segments taken based on linguistic units such as vowels, phonemes, words, and phrases. However, when dealing with automatic and real-time processing, an automatic speech recognizer (ASR) is needed to make use of an utterance-based approach. Although the use of ASRs typically does not degrade the performance of a system [109], [110], it does increase the computational complexity of the system and could represent a threat to users' privacy. On the other hand, a windowing-based approach makes use of a window of time (in milliseconds or seconds), windows of speech activity (defined by pauses or silence), and/or windows of voiced or unvoiced signals. Windowing-based approaches are preferred in real-time systems because they are very fast and computationally efficient. However, this efficiency could be compromised when high amounts of memory space are needed to extract features of interest. While very small windows of time may not provide enough information to determine a change in a behavioral state, longer windows of time provide information similar to the one obtained from utterance-based approaches. This is because, in general, an utterance is comprised of pauses or breath segments and voiced-unvoiced speech segments [111]. Thus, accumulating data from audio frames creates a larger window of speech activity with speech and salient segments, similar to the information of utterance-based approaches. A good balance between performance and computational complexity can be found by evaluating different time window sizes as done in [112]. Here, we discuss works that make use of both approaches with the goal of extracting general information about relevant features.

Before extracting acoustic features, it is common to pre-process the audio signal using a pre-emphasis filter and a window function (e.g., Hamming window) applied to each frame to reduce signal discontinuity and avoid spectral leakage. Table III describes all the identified acoustic features used in the reviewed literature. Acoustic features were grouped by several feature categories: prosodic in speech, conversational characteristics, voice quality characteristics, cepstral coefficients, formant characteristics, frequency spectrum coefficients, and others. The definition of some of the features vary depending on the applied segmentation approach, therefore, we do not define them in here, but information can be found in the references listed in Table III. Because different audio features have been reported to contribute in different ways to the recognition of human behaviors, it is valuable to look more deeply into the level of contribution that various audio features provide toward behavior recognition.

1) Emotions: Lee and Narayanan [110] evaluated, using a feature selection method, a set of prosodic (voice pitch, energy, speech duration, and their statistics) and formant (concentration of acoustic energy at particular frequencies) features extracted at an utterance level to improve the recognition of two emotion classes: negative and non-negative emotions (valence dimension). The feature selection method consisted of evaluating classification accuracies using the k-nearest neighborhood classifier with a leave-one-out cross-validation method. While the authors separated speech data by gender binarism (female, male), the ratio of duration of voiced and unvoiced region, energy median, and F0 (voice pitch) regression coefficient were included in the five-best features for both genders. In this same line, Tahon *et al.* [113] employed an ANOVA test and a classifier to study the contribution of prosodic, cepstral, and voice quality features in the detection of positive and negative emotions (valence dimension). They concluded that the mean and standard deviation (Std Dev) of the relaxation coefficient (a parameter associated with how relaxed is the human voice), the harmonics-to-noise ratio

TABLE III
AUDIO FEATURES FOUND IN THE REVIEWED LITERATURE ASSOCIATED WITH HUMAN BEHAVIOR EFFECTOR CLASSES

Features	*E*	*PF*	*SI*
Prosodic features: Volume amplitude (statistics*), intensity (statistics*), energy (entropy, RMS, linear regression, statistics*), voice pitch (linear regression, statistics*), autocorrelation (maximum peaks, # of peaks), voiced time	[23], [91], [156], [157], [159], [162], [166], [167], [169], [177], [194], [93], [101], [110]–[114], [154]	[79], [117], [119], [120], [122], [123], [163]	[79], [81], [164], [213], [82], [85], [124]–[128], [130]
Conversational features: Turn duration, # of turns, speaking duration (statistics*), speaking rate, overlapping speech duration, interruptions, pause duration (statistics*), # of pauses	[110], [114], [157]	[79], [116], [117], [119]–[123]	[79], [83], [124], [125], [127], [128], [130], [164], [213], [225]
Voice quality features: Zero-crossing rate, harmonics-to-noise ratio (HNR), jitter, shimmer, glottal features (# of glottal pulses, relaxation coefficient (Rd), functions of phase-distortion (FPD))	[23], [91], [177], [111]–[113], [156], [157], [159], [166], [167]	[119]	-
Cepstral features: Shifted delta cepstrum (SDC), mel-frequency cepstral coefficients (MFCC), perceptual linear prediction (PLP) cepstral coefficients, linear prediction-based cepstral coefficients (LPCC), plus their delta and acceleration values	[23], [91], [161], [162], [166], [167], [169], [177], [194], [111]–[113], [154], [156], [157], [159], [160]	-	-
Formant features: Formant frequencies (first and second), bandwidths (first and second), statistics*	[23], [93], [101], [110], [111], [166], [167], [194]	[119], [163]	[213]
Frequency spectrum coefficients: Brightness, center of gravity, distance between the 10 and 90 % frequency quantile, slope between the strongest and the weakest frequency, linear regression, spectral energy (statistics*)	[91], [93], [194], [101], [112], [154], [157], [159], [166], [169], [177]	[119]	[213]
Others: Wavelet coefficients, air pressure distribution in the vocal tract	[115], [162]	-	-

Note. * Statistics include mean, Std Dev, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

(HNR), and unvoiced ratio are of interest for valence detection. Also, of interest resulted a combination of features consisting of the functions of phase-distortion (FPD) (a distortion of the phase spectrum around its linear phase component), voice pitch, energy, and shimmer features.

In the recognition of discrete emotions, the recognition of frustration and calmness can be found. Ang *et al.* [114] showed, through the use of "a brute-force iterative feature selection algorithm", how prosodic features extracted at the utterance level contributed to the recognition of frustration. They concluded that longer durations of vowels or phonemes in an utterance (a word in their case) and slower speaking rates (number of vowels divided by the duration of the utterance) were associated with frustration. In addition, high values in voice pitch features such as maximum pitch in the longest vowel, the maximum overall pitch, the times that the maximum and minimum pitch occurred, the maximum speaker-normalized pitch rise, and the distance of various pitch statistics from the speaker baseline were all associated with frustration, representing the highest percentage of the total information used by the classifier. Other features that were associated with frustration were speaker-normalized RMS energy features, number of dialog exchanges between user and system, and raised voice.

In addition to the direct use of extracted features, some works have applied principal component analysis (PCA) to reduce the dimensionality of the feature vector used to perform classification. Sahoo and Routray [115] estimated the pressure distribution in the vocal tract, which often results in a minimum of 40 feature values that increase depending on the number of vowels present in a given utterance or window of time. Thus, the authors applied PCA and made use of the first 6 principal components to classify calm and aggressive speech segments.

2) Personality Factors: When monitoring elements of personality factors, and also social interactions, there are two types of features that can be extracted: individual-level features and group-level features. Individual-level features are extracted based on audio signals of a single individual and could include any of the acoustic features listed in Table III. Group-level features are extracted from individual-level features; they describe dynamics of a group of people. Thus, they are typically extracted using a window of time with a size in the order of minutes [116].

Related to personality traits, prosodic features have been found to be important for modeling observed extraversion, emotional stability, and openness to experience. Mairesse *et al.* [117] analyzed how those three aspects of personality were correlated to prosodic features. It was found that the maximum voice pitch and the mean, Std Dev, and maximum values of intensity in decibels (dB) were highly correlated with extraversion. Emotional instability was highly correlated with the voiced time and the minimum and mean values of voice pitch, while openness was correlated to the maximum voice pitch values and voiced time. The authors also showed that prosodic features are very good predictors of extraversion in comparison to other types of non-acoustic features. In this sense, analytical studies have shown that extroverts speak more rapidly, with fewer pauses and hesitations than introverts [118]. Moreover, extraversion has also been associated with high values of voice pitch and higher variations in fundamental frequency, shorter periods of silence, and higher voice quality and intensity. This was confirmed by Vinciarelli *et al.* [119] when studying how acoustic features correlated to personality traits. A high perception of extraversion was associated with a high voice pitch and speaking rate. A high center of mass in the power spectrum and a high spectral tilt were correlated with perceptions of less agreeableness. On the contrary, voices with a peakier power spectrum and tendency to be skewed towards higher frequencies are perceived as more agreeable. Furthermore, the perception of conscientiousness was also affected by the same signal features influencing the perception of high agreeableness, together with the speaking rate (people that talk faster are perceived as more competent).

In the case of neuroticism, a high perception of this personality trait was associated with a high voice pitch and first formant mean values. However, no evidence of correlation between audio features and perceived personality traits was found for openness.

Related to the person perception dimensions, Tusing [120] studied how much the amplitude of the speech signals in dB, the voice pitch, and the speech rate in words per minute (wpm) contribute to the perception of dominance. Through a hierarchical regression analysis, it was concluded that the mean amplitude, the amplitude standard deviation, the average voice pitch, and speech rate were correlated with aspects of dominance. This is particularly interesting because it has been noted that dominant people tend to be verbally active while non-dominant individuals are less so. One of the greatest advantages of using speaking rate and features like speaking length [121], [122] to infer dominance revolves on its fast computation and easy use in real-time human behavior monitoring systems. These features were employed by Eagle and Pentland [79], who made use of conversation features such as speaking rate, energy, duration of time holding the floor, interruptions, and turn-taking transition probabilities to build over time profiles of participants' typical social behavior. This allows to recognize relationships and dominant behaviors. In a work by Jayagopi *et al.* [122], features such as speaking turn duration histogram, total successful interruptions, total speaking turns, and total speaking energy also proved to be a good combination of features to identify the most and the least dominant individuals in an interaction. Similar features were shown in [123] to help identify emergent leaders.

3) Social Interactions: Using individual-level features, Hillard *et al.* [124] studied the automatic detection of agreements and disagreements using prosodic and linguistic features. There it was found that prosodic features such as the average, maximum and initial pause duration, the maximum and average voice pitch values, and the average and maximum duration of an utterance are almost as good as linguistic features in identifying segments of agreements. Investigating the automatic detection of the level of interest and involvement of individuals in an interaction, Gatica-Perez *et al.* [125] found through a feature selection method that speech energy, speaking rate, and voice pitch were the best audio features for the task. Moreover, voice pitch values have also been associated with the detection of emphasis during meetings [126]. On the other hand, Cerekovic *et al.* [127] studied the correlation of acoustic features with self-reported and judged evaluations of rapport between a subject and a virtual agent. It was found that interactions with less and shorter pauses, longer speech segments, and louder speech were correlated with high rapport. Turn-taking patterns were also correlated with rapport.

Using group-level features, conversation dynamics have been very well explored. It has been studied that global features such as group speaking interruption-to-turns ratio and the distribution of group speaking turns have been found to discriminate with high accuracy between a competitive meeting and a cooperative meeting [128]. Features such as turn-taking have also been used to identify conversations between two individuals by calculating the mutual information between the turn-taking features of the individual's audio streams [83] and to detect conflicts [129]. Other features such as the sum of all the individual's pause duration, the maximum speaking rate during overlapping speech among individuals, the minimum average turn length among individuals, the total time that at least two people are speaking at the same time (total overlap time), the average energy that is observed for any participant when they are speaking at the same time as at least one other person, and the speaking rate during overlapping speech were reported to have high values in high-cohesion meetings [130].

B. Electrodermal Activity (EDA) Signals

Electrodermal activity (EDA), also known as galvanic skin response, are signals that represent the flow of current between two points of skin contact at which an electrical potential is applied. EDA signals represent properties of the skin that are regulated by changes in sweat glands' secretion, which are controlled by the sympathetic nervous system; sweat secretion increases with increments in emotional arousal. As a result, EDA is considered a good indicator of emotional arousal [131]. EDA signals can be sampled at a rate as low as 4 Hz.

The EDA signal is a time series signal with two activity components, called phasic and tonic, with frequency components of interest between 0.05 and 3 Hz. The tonic component is a slow changing signal, on the scale of tens of seconds to minutes, that is also known as the skin conductance level (SCL). On the other hand, the phasic component, also known as the skin conductance response (SCR), is typically the component considered in human behavior recognition tasks. EDA signals are usually pre-processed to identify and remove movement and respiratory artifacts [132], [133].

Similar to audio signals, in the automatic processing of EDA signals, windows of time are used to extract features of interest. Because EDA signals are slower changing signals than audio signals, the window size used to extract EDA features can vary from 5 seconds to 1 minute. Table IV describes all the identified EDA features used in the reviewed literature. EDA features were grouped by feature categories: raw EDA features, SCR features, SCL features, frequency features, and coupling indexes. General information on their definitions can be found in the references listed in Table IV. In unimodal systems specifically, EDA signals have been used in the recognition of personality factors and aspects of social interactions; and have been consistently processed using coupling indexes.

1) Personality Factors: Empathy has been one of the personality factors monitored using EDA signals. Slovák *et al.* [134] studied the monitoring of empathy in dyads. Raw EDA signals were first smoothed using a rectangular smoothing algorithm and then uniformly scaled based on a running minimum and maximum value taken from each participant from which data was collected. Using a 15-second window with a moving rate of 1 second, signals from pairs of individuals were combined using a Pearson correlation algorithm. In addition, the single session index (SSI), which "represents an index of synchrony over a longer period of time and is calculated as the natural logarithm of the ratio of the sum of positive synchrony divided by the sum of negative synchrony over the specified time," [134] was then computed for the entire recording section (4 minutes). It

TABLE IV
EDA FEATURES FOUND IN THE REVIEWED LITERATURE ASSOCIATED WITH HUMAN BEHAVIOR EFFECTOR CLASSES

Features	*E*	*PF*	*SI*
Raw EDA features: # of local minima, # of local maxima, derivatives, non-stationary index & statistics*	[158], [167], [177]	-	[135]
SCR features: # of peaks, peak amplitude, rise time, recovery time, peak duration, zero-crossing rate of slow response (0-2.4 Hz), & statistics*	[96], [154], [156], [158], [167], [169], [177], [197]	-	[137]
SCL features: Zero-crossing of very slow response (0-0.2 Hz) & statistics*	[154], [167], [169], [177], [197]	-	-
Frequency features: Spectral power coefficients & statistics*	[158], [167], [177]	-	-
Coupling indexes: Pearson's correlation coefficient (PCC), signal matching, instantaneous derivative matching (IDM), directional agreement (DA), Fisher's z-transform of the PCC, single session index (SSI)	-	[134]	[135]–[137], [155], [210]

Note. * Statistics include mean, Std Dev, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

was concluded that high emotional engagement of individuals in the conversation was consistently associated with high EDA synchrony. On the other hand, low emotional engagement was associated with moments of inconsistence or fluctuating EDA synchrony.

2) Social Interactions: In addition to Pearson's correlation coefficient (PCC), other physiological coupling indices that have been found in the literature are signal matching, instantaneous derivative matching (IDM), directional agreement (DA), and the Fisher's z-transform of the PCC. In the area of collaboration, a regression analysis showed that out of the five coupling indices, IDM and DA were good predictors of collaborative behavior [135]. Haataja *et al.* [136] presented an analysis of synchronicity that, first, calculates the average slope of an EDA signal in a 5-second window and then calculates the PCC between EDA signals of two individuals using a moving 15-second window. Similar to the case of empathy, the SSI was calculated but using a window of 2 minutes. Results indicated that physiological synchrony does occur during collaborative learning at a statistically significant level. Because the analysis was performed offline, found moments of synchrony could not be correlated to specific monitoring instances. However, results do suggest that physiological synchrony might be a relevant condition when joint understanding is better built within groups. In an effort for studying the dynamics of collaboration related to the degree of physiological activation of triads, Pijeira-Díaz *et al.* [137] calculated the number of peaks per minute in SCR signals using a moving window with a window width of 1 minute and a moving step of 250 ms, and then calculated the arousal DA as a measure of the synchrony degree. Results showed that most of the time participants were at different arousal levels, but when they were in synchrony it was mostly in the low arousal level. Although results were not correlated with specific instances, authors showed the potential of using arousal DA to characterize collaborative behaviors.

TABLE V
EEG FEATURES FOUND IN THE REVIEWED LITERATURE ASSOCIATED WITH HUMAN BEHAVIOR EFFECTOR CLASSES

Features	*E*	*SI*
Time domain features: Power, derivatives, Hjorth features (activity, mobility, complexity), non-stationary index, fractal dimension, higher order crossings (HOC), & statistics*	[96], [145], [226]	-
Frequency domain features (per band): Energy spectrum (ES), power spectrum, power spectral density (PSD), differential entropy (DE), rational asymmetry (RASM) of DE features in a channel pair, differential asymmetry (DASM) of DE features in a channel pair, differential caudality (DCAU) between DE features, higher order spectra (HOS), & statistics*	[144]–[146], [158]	[170]
Time-frequency domain features: Hilbert-Huang spectrum (HHS), discrete wavelet coefficients (DWC)	[145]	-

Note. * Statistics include mean, Std Dev, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

C. Electroencephalography (EEG) Signals

Electroencephalography (EEG) signals represent the electrical activity of the brain. Systems that record EEG signals can have as few as one electrode channel to as many as 256 channels. The placement of EEG electrodes along the scalp is of great importance. Thus, their placement adheres to international standards such as the 10/20 system (also known as International 10/20 system) [138], 10/10, or 10/5 systems [139], the last two also known as the Modified Combinatorial Nomenclature (MCN). These standards aim to standardize the exact position of each electrode and assign names to each of them to facilitate the identification of the brainwave location that may serve a specific brain function. For example, in the area of emotion recognition, specific electrode positions are of interest. T3 and T4, electrodes placed in the temporal lobe regions, are found to be near emotional processors. P3, P4, and Pz, electrodes placed in the parietal brain region, are located near sources that reflect activities of perception and differentiation. While frontal lobe electrodes (i.e., F3, F4, F7, F8) have proximity to sources of emotional impulses and have been used for emotion recognition [140], [141]. EEG signals are typically sampled at a rate of $\sim$256 Hz but can be sampled at a lower rate depending on the signal components of interest.

As EEG signals have a low signal-to-noise ratio and are prone to muscle movement artifacts [142], pre-processing of these signals includes filtering and signal inspection for artifact removal. Before extracting features, typically a window function (e.g., Hamming window) is applied to each window of time or frame of data to reduce signal discontinuity and avoid spectral leakage. The length of these time windows is at least 1 second. Table V describes all the identified EEG features used in the monitoring of human behavior. EEG features were grouped by feature categories: time domain features, frequency domain features, and time-frequency domain features. General information on their definitions can be found in the references

listed in Table V. Traditionally, EEG signals have been analyzed using event-related potential (ERP) features. However, when EEG signals are analyzed based on identified ERPs, an event (or trigger) needs to be identified, and then features describing the response to that event are extracted [143]. This approach is not suitable for real-time implementation since it is unknown when an "event" will happen. On the other hand, when EEG signals are analyzed using either time, frequency, or time-frequency domain features, EEG signals are first divided into frequency bands containing slow, moderate, and fast brainwaves that are associated with specific brain states (e.g., sleep, relaxed, and alert). These frequency bands are delta band (1–4 Hz), theta band (5–8 Hz), alpha band (9–12 Hz), beta band (13–25 Hz), and gamma band (>25 Hz). However, the exact frequency values used to extract the frequency band can vary across researchers by 1 or 2 units of Hz per band. Typically, features are extracted specifically per frequency band. In unimodal systems, EEG signals have been used mostly for the recognition of individual emotions.

Duan *et al.* [144] extracted frequency domain features in five frequency bands from signals recorded from a 62-channel electrode cap to classify positive or negative emotional states of the individuals participating in their study. All features used were smoothed using a linear dynamic system (LDS) approach. They found that emotional states relate to EEG signals in the gamma band more closely than other frequency bands and that using differential entropy (DE) as a feature provides better results than using more traditional features such as energy spectrum (ES). Likewise, Jenke *et al.* [145] evaluated different time, frequency, and time-frequency feature sets from signals recorded from a 64-channel electrode cap. Using feature selection methods, it was concluded that features such as power spectrum, higher order spectra (HOS), Hilbert-Huang spectrum (HHS), and, discrete wavelet coefficients (DWC) computed from beta and gamma bands were better at classifying emotions. Zheng *et al.* [146] investigated not just the frequency domain features and critical frequency bands for the recognition of three emotion states (positive, neutral and negative), but also the performance of a combination of four, six, nine, and 12 channels in the recognition of the three emotions. The authors concluded that DE performed better as a feature when compared to power spectral density (PSD), differential asymmetry (DASM), rational asymmetry (RASM), and differential caudality (DCAU). In addition, as noted in previously discussed works, they also confirmed that beta and gamma oscillatory brain signals are more related to emotion processing than other frequency bands. Using a weight distribution of a trained deep belief network (DBN), the 12 channels that collect the most emotional information are FT7, FT8, T7, T8, C5, C6, TP7, TP8, CP5, CP6, P7, and P8 (named based on the MCN system). If reduced to four channels, they found them to be FT7, FT8, T7, T8, wherein the 10/20 system, T7 and T8 are T3 and T4, respectively.

D. Electrocardiogram (ECG) Signals

Electrocardiogram (ECG) signals represent the electrical activity of the heart. Frequency components of interest in ECG signals are below the 20 Hz, although a commonly used sampling frequency is of 1 kHz. A heartbeat (or cardiac cycle) is associated with ECG signal phases and specific signal characteristics. A complete cardiac cycle is made up of five waves that construct an ECG signal, namely P wave, Q wave, R wave, S wave, and T wave. From those five waves, five signal phases are identified: PR interval, PR segment, QRS complex, ST segment, and QT interval. Each of them is associated with how the electrical signal travels through the heart. For the heart rate measurement (or frequency of the cardiac cycle), the QRS complex is the most important signal phase because the heart rate is calculated from the time between any two consecutive QRS complexes (R-R interval).

TABLE VI
ECG FEATURES FOUND IN THE REVIEWED LITERATURE ASSOCIATED WITH HUMAN BEHAVIOR EFFECTOR CLASSES

Features	*E*	*SI*
Time domain features: Heart rate (HR) (expressed in beats per minute (bpm)), inter-beat interval (IBI) (measured in ms), zero-crossing rate, non-stationary index, heart rate variability (HRV), & statistics*	[152], [156], [158], [167], [169], [177]	-
Frequency domain features: Spectral power, power spectral density, spectral entropy, derivatives & statistics*	[158], [167], [177]	-
Coupling indexes: Pearson's correlation coefficient (PCC), Fisher's z-transform of the PCC, weighted coherence	-	[155], [210]

Note. * Statistics include mean, Std Dev, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

Similar to other physiological signals, ECG signals are prone to noise and artifacts, which are typically tackled at the input of the signal acquisition system [147] or at the pre-processing stage. A review of this topic can be found in [148]. Noise and artifact removal of ECG signals is important before feature extraction. Table VI describes all the identified ECG features used in the monitoring of human behavior. ECG features were grouped by feature categories: time domain features, frequency domain features, and coupling indexes. General information on their definitions can be found in the references listed in Table VI. In unimodal systems, ECG signals are used to monitor individual's emotional arousal states through parameters such as heart rate (HR) (expressed in beats per minute (bpm)), inter-beat interval (IBI) (measured in ms), and heart rate variability (HRV) [149]–[151]. In addition, Quintana *et al.* [152] used correlation analysis to study how different social conditions affect HRV and its relation to emotional states. They concluded that high levels of HRV during resting state are associated with improved emotion perception, while reduced HRV is associated with impairments in social cognition.

E. Multi-Signal Modalities

Signals from multi-sensor modalities have been used to increase the robustness of human behavior monitoring systems. However, integration of multiple sensors involves managing inconsistencies in the collected data before feature extraction.

TABLE VII
SENSOR SIGNAL FEATURES USED IN MULTIMODAL SYSTEMS

Features per sensor modality	*E*	*SI*
RF sensor: Raw received signal strength indicator (RSSI) values, duration in time of a RSSI value, mean of measurements from two RSSI RF signals, difference between two RSSI RF signals	[199], [200]	[76], [165]
IR sensor: Number of detected encounters with another IR sensor, sum of lengths of all encounters, and length of an encounter	-	[81], [82], [85]
Accelerometer, gyroscope, and magnetometer: Signal energy, energy-entropy, correlation coefficient between axis, pitch, roll, peak value in frequency domain, statistics*	[91]–[94], [97], [101]	[81], [82], [85], [165]
Skin temperature: Derivatives, spectral power in low frequency bands, PCC, weighted coherence, statistics*	[154], [156], [158], [197]	[155]
Respiration: Signal energy, derivatives, breathing rhythm, breathing rate, sub-band power spectral, PCC, weighted coherence, statistics*	[154], [156], [158]	[155]
Blood volume pulse: Mean signal, variance, sub-band power spectral, power spectral density, heart rate, heart rate variability, blood flow, pulse, statistics*	[154], [156], [197]	-
Electromyogram: Statistics*	[96], [154]	-
Eye-tracker: Pupil diameter, gaze distance, eye blinking, gaze coordinates, statistics, coupling indexes	[158]	[155]
EOG: Blink rate, blink amplitude, power of blink amplitude, statistics*	-	[170]

Note. * Statistics include mean, Std Dev, variance, skewness, kurtosis, slope, median, maximum, minimum, range.

Different sensor signals are typically collected using different sampling frequencies, they use different pre-processing methods, and they require different windows of time to extract features. All of these contribute to inconsistencies in the data collected across sensor modalities and present a great challenge for data synchronization, which is important to achieve robustness in human behavior monitoring systems. Nonetheless, when signal features from two or more sensing modalities are used, the reviewed literature identifies two common methods to combine information: feature-level fusion and decision-level fusion. In feature-level fusion, the features extracted from individual sensors are consolidated into a single feature set. A simple solution to synchronize extracted features at the feature-level fusion is to extract them using the largest window size among the selected sensor modalities and then build a single feature vector. Thus, statistics are commonly employed in the feature extraction process. In decision-level fusion, also called model-level, the decisions from multiple classifiers (usually one classifier per sensor modality) are combined into a common decision. More on the theory of fusion mechanisms can be found in [153]. As performed in the discussion of features from audio, EDA, EEG, and ECG signals, we focus on discussing works performing feature-level fusion and the correlated or best-performing set of features from combined sensor modalities. Table VII lists additional sensing modalities used in the review literature together with the type of features that are typically extracted from each of them.

1) Emotions: In [154], a total of five sensors were used for the recognition of four emotions. Features from audio, EDA, EMG, PPG, skin temperature, and RSP signals were extracted. Through a sequential backward selection algorithm, features such as the sub-band spectral entropy from PPG, the number of peaks within 4 seconds in EDA and EMG, and the mean values of the MFCCs in the speech features stood out in the recognition of the four emotions. On the other hand, [91] and [93] made use of audio and movement (from accelerometers and gyroscopes) signals to recognize anxiety levels and other individuals' well-being characteristics, respectively. Both made use of a Pearson product-moment correlation coefficient (PPMCC) analysis to investigate the most relevant features associated with anxiety and well-being. In [91], it was found that at least brightness and MFCC5 from speech, and Std Dev of the axis of gyroscopes and their peak value in frequency domain were highly correlated with the degree of anxiety of the individuals in the study. Likewise, [93] found that the formants, energy, entropy, and brightness features from audio signals and both time and frequency domain features from accelerometers and gyroscope were strongly correlated with mental health questionnaire responses.

2) Social Interactions: Gips and Pentland [81] and Laibowitz *et al.* [82] used three sensors for the recognition of interest during a social encounter. Initially, a 15-dimensional feature vector was constructed per dyad encounter with features from accelerometers, microphones, and IR sensors. Based on a correlation analysis, the six highest ranked encounter features for the recognition of interest was: Std Dev of accelerometer measurements in the x-axis and y-axis, mean and Std Dev of average audio signal amplitude, mean average audio difference between averaged readings, and Std Dev of the difference between the average amplitude and the average difference. In the use of combination of physiological signals, Pun *et al.* [155] used a total of five sensors for the recognition of collaborative behaviors. Coupling features from EDA, ECG, eye-tracker, skin temperature, and RSP signals were extracted. Through a fast-correlation based filter with mean squared linear regression, the correlation between the extracted features and the degree of perceived collaboration was determined. Coupling features were calculated using the signals from dyads in an interaction. From the physiological signals, the coherence of the IBI in the very low frequencies (0.003–0.05 Hz) and the in the low frequencies (0.05–0.15 Hz) were correlated with aspects of collaboration. While from eye-movement signals, the number of times participants looked at the same place at the same time and the number of times participants looked at the same place within a ± 6 second window were correlated with collaborative behaviors. Related to group cohesion, Zhang *et al.* [85] made use of a wearable sociometer badge with accelerometer, microphone, and an IR sensor to measure cohesion at an individual level and at a group level. Using Pearson correlation coefficients, it was found that at the individual level, the mean movement energy was positively correlated with cohesion task. At a dyadic level, the correlation of the vocal activities was also positively correlated with cohesion task.

F. Analysis and Discussion

To eliminate redundant information and optimize algorithms for real-time implementation, it is important to perform correlation analysis or feature selection to analyze the contribution of signal features in the recognition of a human behavior effector. As noted in Tables III–VII, a wide range of features from different sensor modalities have been employed for the recognition of human behavior effectors. Although not all works referenced in the tables performed correlation analysis or feature selection on extracted signal features, significant consistency exists among the best-found features to be used in recognizing emotions and those to be used in recognizing personality factors and social interactions.

From the agglomeration of references in Tables III–VII, one can observe that the most common features for recognizing emotions are: prosodic, cepstral, voice quality, and frequency spectrum coefficients from audio signals; SRC features from EDA signals; frequency domain features per frequency band from EEG signals; and time domain features from ECG. More specifically, from prosodic features of audio signals, features related to voice pitch appear to greatly contribute to the recognition of positive and negatives emotions (i.e., emotional valence levels). From EEG signals, the DE feature extracted from the gamma frequency band has also proven to be effective in the recognition of positive and negative emotions. Moreover, from ECG signals, the HRV, which can also be determined from PPG signals, has been found to be a good indicator of emotional valence, emotional arousal and emotion perception. On the other hand, features from sensor signals used in multi-signal modalities such as Std Dev of gyroscope's axis values and their peak value in frequency domain have been found to be correlated with anxiety levels.

In the case of personality factors and social interactions, from audio signals, prosodic and conversational features are the most commonly used. More specifically, from prosodic features, voice pitch has proven to greatly contribute to the recognition of extraversion, dominance, and emphasis during meetings. On the other hand, conversational features such as speaking rate and speaking length have proven to contribute to recognizing cooperative meetings, in addition to extraversion and dominance. In general, speaking length and speaking rate are also attractive for real-time use because of their low computational complexity and fast computation. For social interactions alone, other commonly used features found to be relevant in the recognition of social interaction elements, such as collaboration and cohesion, are coupling indexes from EDA signals; distance between individuals and duration of the encounter obtained from IR sensor signals; eye-movement related features from eye-tracker sensor signal; and Std Dev features of accelerometer measurements in the x-axis and y-axis.

To date, analyses of features' contribution in the recognition of human behavior effectors come from works on unimodal systems and less so from works in multimodal sensor systems. This could, arguably, be due to the large number of works in unimodal sensor systems. Still, from observation, the most common sensor signals' combinations used in multi-sensor modalities include microphones with physiological sensors and/or movement and proximity sensors, and combinations of physiological sensors. However, further research is encouraged in the evaluation of the best feature or features to be used in multi-sensor modalities for the recognition of human behavior effectors. As sensor features are identified as contributing to the recognition of more than one human behavior effector, more optimized and robust systems could be designed. For example, voice pitch, from audio signals, has been observed to be a good contributor for the recognition of all human behavior effectors. Thus, using voice pitch when designing a system to recognize multiple human behavior effectors could help increase system efficiency.

Despite the fact that classification model performance is impacted by the quality of features and outliers in the data, many papers fail to analyze the statistical distribution of the signal features. This analysis is necessary to select appropriate (best performing) classification models, and researchers are highly encouraged to make this part of their standard design process.

VI. Computational Models for Human Behavior Recognition

Based upon the features extracted from sensor signals, computational models are trained and used to predict or classify human behavior. Therefore, the performance of computational models can depend on the set of features provided. Likewise, the effectiveness of signal features can also depend, in part, on the type of computational method used to evaluate the features contribution.

The two principal types of computational models employed in the human behavior recognition literature are classification and regression models. Classification models focus on recognizing discrete or categorical classes, while regression models focus on predicting continuous numerical values. The use of a computational model is application dependable. For example, the problem of emotion recognition can be treated as one with categorical values (e.g., happy, sad, neutral) or as one with continues numerical values (i.e., reflecting levels of arousal and valence based on a numerical scale).

An analysis of reported computational methods used in the monitoring of human behaviors was performed as follows. First, reviewed literature was grouped based on their use of classification and regression models. Then, within each of the two model groups, different types of models and the number of predicted or classified classes were summarized to illustrate the design space employed in the literature. This summary analysis allowed us to make observations regarding the most commonly used computational models, which are presented at the end of the section. Specifically related to classification models, we analyzed and compared their accuracy values to define the state-of-the-art system performances that may help drive the future design of real-time human behavior monitoring systems.

A. Classification Models

In general, based on the reviewed literature, classification models have been widely used in emotion, personality factors,

TABLE VIII
CLASSIFICATION MODELS FOUND IN THE REVIEWED LITERATURE ASSOCIATED WITH HUMAN BEHAVIOR EFFECTOR CLASSES

Models	*E*	*PF*	*SI*
Support Vector Machine (SVM): Classic SVM, adaptive SVM, and incremental SVM	[23], [96], [111], [144], [146], [156], [158]–[160]	[122], [163]	[128], [130]
k-Nearest Neighbor (k-NN)	[23], [110], [144], [146], [162]	-	-
Naïve Bayes (NB)	[23], [96], [145], [157]	-	[130]
Log-likelihood ratio	-	-	[128]
Logistic regression	[146]	[163]	[85]
Linear regression	-	-	[82]
Linear Discriminant Analysis (LDA)	[110], [154]	-	-
Decision and Regression Tree	[96]	-	[124], [165]
Random Forest	[156]	-	-
Hidden Markov Models (HMMs)	[115]	-	[125], [164]
Gaussian Mixture Model (GMM)	[111], [162]	-	-
Neural networks: Convolutional NN, Multilayer perceptron (MLP), self-organizing map, deep belief networks (DBNs)	[111], [112], [146], [161]	-	-
Partial Least Squares-Discriminatory Analysis (PLS-DA)	[23]	-	-
Latent Dirichlet Allocation model	-	[116]	-
Sets of rules: Rule-based, rank-level fusion, collective classification approach	-	[123]	-
Clustering models: k-means	[91]	-	-

and social interaction recognition tasks. The reviewed literature presents variations in the number and type of classes that classification models are trained to recognize and variations in the classification models being employed. A summary of the classification models employed in the reviewed literature associated with human behavior effector classes can be found in Table VIII.

1) Emotions: Lee *et al.* [110] investigated the performance of a k-Nearest Neighbor (k-NN) and a Linear Discriminant Analysis (LDA) classifier to predict two emotion classes (negative and non-negative) when using audio data from male and female subjects, separately. While for female data, LDA consistently performed better than k-NN, for male data there were cases in which k-NN performed better than LDA. Gu *et al.* [91] made use of a K-means classifier to recognize high anxiety and low anxiety using features from audio signals. The authors obtained 72.73% of performance accuracy by using just two features: brightness and MFCC. In this line, Sahoo and Routray [115] trained Hidden Markov Models (HMMs) to detect aggression and calmness also using audio signals. By using pressure distribution features a performance accuracy of 93.5% was achieved. Later, by using the same features, the authors trained an HMM to recognize four emotion classes (anger, boredom, happiness, and neutral) achieving an 80% overall recognition accuracy. On the other hand, using EEG signals, Duan *et al.* [144] evaluated two classifiers, a Support Vector Machine (SVM) and a k-NN to predict two emotion classes (positive and negative emotion). In general, SVM outperformed k-NN achieving a performance accuracy of up to 86.69%. Using a multimodal sensor system, Chanel *et al.* [156] investigated the performance of Random Forest and SVM classifiers in predicting emotional and non-emotional moments using audio and physiological (EDA, ECG, BVP, skin temperature, and respiration) signals during a social interaction. The authors investigated the performance of decision-level fusion by combining the output scores of classifiers trained on signal features from each individual in the interaction. Regardless of the classifier type (Random Forest or SVM), it was found that by adding emotional information from all individuals in the interaction, the emotional response of one individual can be predicted with higher accuracy than just using the classification model from the individual of interest.

Related to the recognition of three emotion classes, Zheng and Lu [146] investigated the performance of four classifiers in predicting positive, neutral, and negative emotion classes using EEG signals. The four classifiers were deep belief networks (DBNs), SVM, logistic regression, and k-NN with resulting average classification accuracies of 86.08%, 83.99%, 82.70%, and 72.60%. However, the highest reported accuracy of DBNs was by taking EEG features from 62 channels, whereas the highest reported accuracy of SVM was 86.65% when taking EEG features from 12 channels.

Related to the recognition of four emotion classes, Kim [154] trained a LDA classifier in combination with a sequential backward selection to predict low and high arousal and high and low valence using audio and physiological (EDA, ECG, BVP, EMG, skin temperature, and respiration) signals. The author trained a model for each subject (three in total) and a subject-independent model achieving an average accuracy of 78.67% and 55%, respectively. Similarly, Vogt *et al.* [157] trained a Naïve Bayes (NB) classifier to predict four emotion classes (joy, satisfaction, anger, and frustration) but just using audio signals. The authors trained subject-dependent models for 29 subjects, achieving accuracy values that ranged from 24% to 74%, with an average of 55%. They also trained a subject-independent model using data from 10 subjects achieving a 41% recognition accuracy. Their use of NB was motivated by its fast computation and ability to take high dimensional feature vectors. However, Vogt *et al.* suggested that a more accurate classifier would be an SVM and that with a vector size under 100 features, it could be suitable for real-time implementation. In this line, using EEG and eye gaze signals, Soleymani *et al.* [158] trained SVM subject-dependent models to predict four emotion classes (high and low arousal and high and low valence). Classification accuracies for arousal and valence were 67.7% and 76.1%, respectively. Using audio signals, Abdelwahab and Buso [159] investigated the use of two modified versions of SVM to classify the same four emotion classes (high and low arousal and high and low valence). They trained an adaptive SVM model and an incremental SVM model, which aims at maintaining or improving their classification performance even under mismatched training and testing conditions. The authors concluded that both methods provide similar performance, but a precise accuracy value was not reported. On the other hand, Wu and Liang [111], also using audio signals, trained three types of models, Gaussian Mixture Model (GMM), SVM, and

a multilayer perceptron (MLP) to predict four emotion classes (neutral, happy, angry, and sad). A Meta Decision Tree (MDT) was then used for classifier fusion, achieving an overall performance accuracy of 80%. However, the results from SVM alone were close to the results of MDT fusion classifier because the MDT is a classifier selection approach instead of a combination of all classifiers. Moreover, Cen *et al.* [160] trained a SVM model for offline and real-time recognition of the same four emotional states (neutral, happy, angry, and sad) also using just audio signals. Their results showed a 90% and 78.78% classification accuracy for automatic offline and real-time emotion recognition, respectively. In addition, Girardi *et al.* [96] investigated the performance of SVM, J48 (algorithm based on decision trees), and Naïve Bayes (NB) on predicting high and low arousal and high and low valence by using physiological signals such as EDA, EEG, and EMG. Results showed that SVM outperforms the other classifiers and that EEG signal features alone provided the best performance accuracy for valence classification, while EEG+EDA performed the best for arousal classification. On the other hand, using a Convolutional Neural Network (CNN) to predict these same four emotion classes, Rajak and Mall [161] using audio signals, specifically, MFCC features achieved a classification accuracy of 76.2%.

In the recognition of more than four emotion classes, Jenke at al. [145] trained NB subject-dependent models using EEG signals to predict five emotion classes (happy, curious, angry, sad, and quite) achieving a performance accuracy of 36.80%. Later, Lanjewar *et al.* [162] made a comparison between the performance of a GMM and a k-NN to predict six emotion categories using audio signals. In general, their results showed that the GMM performed better than the k-NN model with a 66% and 52% of classification accuracy, respectively. However, the speed of computation is faster for the k-NN classifier than for GMM, which makes it attractive when time constraints are critical to consider, like for real-time applications [162]. The computational time of GMM increased when the number of features increased in the training phase. However, it was noted that GMM was better at predicting angry and sad emotion classes, while k-NN performed better at predicting happy and angry emotion classes. Also using audio signals, Balti and Elmaghraby [112] implemented a self-organizing map with a response integration approach to predict seven emotion classes (anger, boredom, disgust, anxiety/fear, happiness, sadness, and neutral), achieving a 70.86% performance accuracy. Likewise, Jing *et al.* [23] investigated the performance of SVM, k-NN, NB, and Partial Least Squares-Discriminatory Analysis (PLS-DA) when used to predict seven emotion classes (sad, joy, fear, surprise, neutral, anger, and disgust) by using audio and EGG signals. The authors evaluated the models using acoustic features only and combined feature sets independently for males and females. However, the results consistently showed that SVM got a higher average emotional recognition accuracy for both genders, when compared to the other classification models, with a classification accuracy of $\sim$72%.

2) Personality Factors: Using audio signals, Jayagopi *et al.* [122] trained an unsupervised classification model and an SVM model to predict the most-dominant person and the least-dominant person in a group conversation. The unsupervised model computed either the largest or smallest accumulated value of each extracted feature, depending on whether the goal was to predict the most dominant or the least dominant person. In addition, two SVM models were trained. One to predict the most and the non-most dominant person of the group conversation, and another one to predict the least and the non-least dominant person in the same group conversation. Their results showed that SVM performed better than the unsupervised model in predicting the most-dominant person, being their best performance accuracies 91.2% and 85.3%, respectively. On the other hand, both models performed the same when predicting the least-dominant person with an 83.9% accuracy. The same author, in [116], also using audio signals, trained a Latent Dirichlet Allocation model to predict three classic leadership styles: autocratic, participative, and free-rein, achieving a 79.20% classification accuracy. Likewise, Sanchez-Cortes *et al.* [123] evaluated four approaches using audio signals to infer an emergent leader in a group. The four approaches were a rule-based approach (search for the person with the highest feature value in a group and selects that as the leader), a rank-level fusion (extension of rule-based that handle fusion of multiple features), SVM, and a collective classification approach. Results showed that the rank-level fusion provided the best performance with a 72.5% of accuracy. It also performed the best in identifying perceived dominance with 65% of accuracy. Related to personality traits, Mohammadi and Vinciarelli [163], also using audio signals, evaluated the performance of a logistic regression and an SVM in predicting high and low extraversion, agreeableness, conscientiousness, neuroticism, and openness. Results suggest that logistic regression performs better than SVM in predicting conscientiousness and neuroticism with a 72.55% and 66.10% classification accuracy, respectively. On the other hand, SVM performed better than logistic regression in predicting extraversion, agreeableness, and openness with 73.45%, 63.10%, and 52.75% classification accuracy, respectively.

3) Social Interactions: Similar to the previous sub-sections, most of the literature reported in here has trained their models with features from audio signals. Using prosodic and conversational features, Hillard *et al.* [124] trained a Decision tree (DT) classifier to predict moments of agreement and disagreement during meetings. They achieved an overall performance accuracy of 64%. Similarly, Jayagopi *et al.* [128] evaluated a log-likelihood ratio model and an SVM model in classifying conversational group dynamics into cooperative-type or competitive-type. Using an SVM with a quadratic kernel, 100% classification accuracy was obtained.

In line with meetings, McCowan *et al.* [164] trained a HMM to predict eight meeting actions (monologues from individuals (total of 4), note-taking, presentation, discussion, and white-board talk) achieving an 83.9% classification accuracy. Also using HMM, Gatica-Perez *et al.* [125] predicted two levels of interest, high and low, during a meeting. By training an HMM with a feature vector constructed from calculating the mean of the features from all the subjects in the interaction, an 84%

TABLE IX
REGRESSION MODELS FOUND IN THE REVIEWED LITERATURE ASSOCIATED WITH HUMAN BEHAVIOR EFFECTOR CLASSES

Models	*E*	*SI*
Support Vector Regression (SVR)	[166], [167], [169]	-
Regression Trees	-	[155]
Least Squared regression	-	[155]
Neural networks: Long short-term memory recurrent neural network (LSTM-RNN), Feed-forward (FF), Bilateral long short-term memory (BLSTM)	[166], [167], [169]	-
Structured regression model: Continuous conditional neural field (CCNF), continuous conditional random field (CCRF)	-	[170]

recall and 63% precision performance measures were achieved, while by just concatenating the features from all the subject an 80% recall and 58% precision performances were achieved. Also investigating levels of interest but during social encounters, Laibowitz *et al.* [82] trained a Linear Regression model using accelerometer signals, in addition to audio signals. Their model achieved an 86.2% classification accuracy.

Related to cohesion, Hung and Gatica-Perez [130], evaluated the classification performance of an NB model and an SVM model when predicting high and low cohesion using audio signals. Both classifiers showed similar classification performances, achieving up to 90% accuracy. Moreover, Zhang *et al.* [85] employed a logistic regression classifier to recognize between task cohesion and social cohesion among dyads by using audio, accelerometer, and IR signals. Their approach achieves 80.30% and 64.62% classification accuracy when predicting task cohesion and social cohesion, respectively. On the other hand, Katevas *et al.* [165] used a XGBoost regression tree classifier to detect interactive groups of various sizes (node and group level) by using accelerometer, gyroscope, and RF signals, achieving a 94% performance accuracy.

B. Regression Models

In general, works that have made use of regression models are focused on the prediction of emotions and social interactions. Regression models have been found to be particularly attractive when it is of interest to predict or recognize levels of emotional arousal, emotional valence, collaboration, and vigilance on a continuous numerical scale. A summary of the regression models employed in the reviewed literature associated with human behavior effector classes can be found in Table IX.

1) Emotions: Wöllmer *et al.* [166] introduced a framework for continuous monitoring of arousal and valence levels using audio signals. The authors evaluated two regression models: Support Vector Regression (SVR) and a long short-term memory recurrent neural network (LSTM-RNN). Their results showed that LSTM-RNN performed better than SVR at predicting arousal levels with a Mean Squared Error (MSE) performance measurement of 0.08 and 0.10, respectively. On the other hand, both regression models performed the same at predicting valence levels with a MSE of 0.18. Ringeval *et al.* [167] used a hybrid decision-fusion based on SVR with a lineal kernel and Neural Networks (NN) to predict arousal and valence emotional levels based on data from audio, video, EDA, and ECG sensors obtained from the AV+EC 2015 database [168]. For NN, they explored three types of architectures: feed-forward (FF), LSTM, and bilateral long short-term memory (BLSTM). The authors found that SVR performs best on the audio features for valence prediction with a 0.069 Concordance Correlation Coefficient (CCC) and NN performs best on EDA features for arousal with a 0.79 CCC. Moreover, FF provided the best performance for EDA features. Their hybrid decision-fusion method achieved the best arousal prediction with a 0.228 CCC and 0.173 RMSE performance metric using audio features, while achieved their second-best valence prediction performance with a 0.195 CCC and 0.119 RMSE using EDA features. However, when the authors employed decision-fusion on their multi-modal data, their results improved achieving 0.444 CCC and 0.164 RMSE for arousal prediction, and 0. 382 CCC and 0.113 RMSE on valence prediction, demonstrating the value of a multi-modal approach. Also using SVM and LSTM models, Brady *et al.* [169] used a decision-level approach to predict these arousal and valence levels. The authors trained an SVR model for audio signals and a LSTM for physiological signals (EDA and ECG) and combined their decisions using a Kalman filter framework. They found that models for ECG and EDA provided significant performance improvements for valence prediction, obtaining 0.364 CCC and 0.117 RMSE for models trained with HR and HRV data and 0.177 CCC and 0.124 RMSE for EDA data.

2) Social Interactions: Contrary to emotion recognition, which mainly focuses on predicting arousal and valence levels, in the area of social interactions, the target classes vary greatly from one work to another. Chanel *et al.* [155] used Bag of Regression Trees (BRT) and Least Squared regression with fast-correlation based filter (FCBF LS) to predict collaborative behaviors (e.g. degree of conflict, confrontation, and emotional management) based on data from EDA, ECG, skin temperature, respiration, and eye-tracker. Physiological and eye-tracker data were treated separately, and different regression models performed differently based on the sensor data modality and the targeted collaborative behavior. For example, the FCBF LS model achieved the lowest RMSE value, with a 0.44 RMSE performance value, when using eye-tracker data to predict the degree of convergence in a group of people. However, the BRT model performed better at predicting confrontation using physiological signals when compared to the FCBF LS model achieving a 0.60 RMSE performance value. On the other hand, Zheng and Lu [170] employed an SVR with radial basis function to estimate the level of vigilance based on data from EEG and EOG. The authors introduced a continuous conditional neural field (CCNF) and a continuous conditional random field (CCRF) to the design of their vigilance estimation model with the goal of incorporating the temporal dependency present in vigilance. It was demonstrated that the fusion of multimodal sensor features improves model performance, achieving 0.09 RMSE performance value, compared to features from a single modality that achieved 0.12 and 0.13 RMSE performance values for EOG-based and EEG-based methods, respectively. In

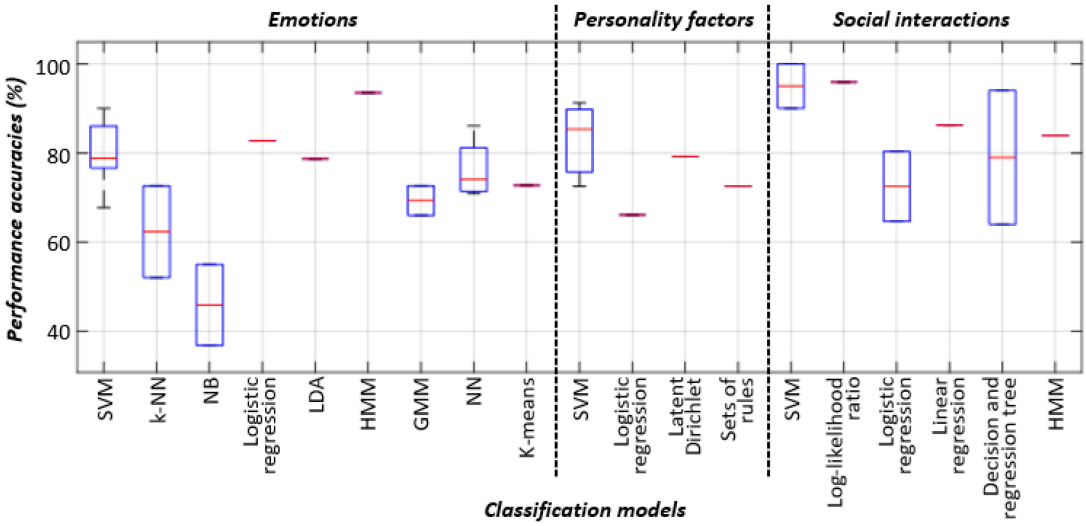

Fig. 9. Summary analysis of reported performance accuracies of classification models per human behavior effector groups. The central mark indicates the median accuracy, and the top and bottom edges of the box indicate the 75th and 25th percentiles, respectively. The whiskers extend to the most extreme accuracy values not considered outliers. These results were obtained by analyzing data from the references in Table VIII.

addition, the temporal dependency-based models demonstrated to also enhance vigilance estimation.

C. Analysis and Discussion

A wide range of computational models, as noted in Tables VIII and IX, have been employed for the recognition of human behavior effectors. To deeply analyze the use of classification models, performance metrics related to the accuracy values reported per classification model are organized by effector class and presented in Fig. 9. From the agglomeration of references in Table VIII, it can be observed that SVM has been the most popular classification model used for the recognition of human behavior effectors followed by k-NN and NB. In addition, from Fig. 9, it can be observed that SVM provides one of the highest levels of accuracy across all effector classes. On the other hand, k-NN and NB have been specifically used in emotion recognition, and although they follow SVM in popularity, their levels of accuracy are of the lowest across all other employed classification models. An important factor to consider when evaluating the performance of classification models is the number of classes that they are trained to predict. For example, in Fig. 9 under emotions, HMM reports the highest accuracy but classifies just two classes, whereas the accuracies reported for SVM are for models trained to recognize from two to four classes. Moreover, the accuracy and complexity of these computational models vary depending on 1) the number of classes that they are trained to predict and 2) the quantity of information (number of features) that they take to accurately predict a class. Both of these factors are also critically important when considering real-time implementations. The four classification models that have been trained to recognize human behaviors in real time are k-means [91], HMM [115], NB [157], and SVM [130], [160].

Our review has identified that classification models have been more widely employed than regression models. This indicates that the problem of identifying human behaviors using sensor technologies has generally been treated as a "discrete problem" rather than a continuous one. However, it has been argued that human behaviors change gradually, in a continuous scale rather than in discrete states [171]. Thus, the use of continuous numerical values for the recognition of such behaviors may be preferred. To date, the use of regression models to treat behavior recognition as a continuous case (i.e., using continuous numerical values to recognize or predict a behavior) has varied with the behavior effector class being monitored; SVR and NN regression models have been common for emotion recognition, while regression trees, least-squared regression, and structured regression models have been used in the prediction of aspects of social interactions. A general observation related to regression models is that, although these models have been employed in the automatic recognition of human behaviors, so far, they do not appear to have been used in real time. However, as regression models are attractive for the prediction of continuous classes, further study of these models for the real-time prediction of human behavior effector classes is highly encouraged. In addition, although there are a limited number of works employing regression models to predict a behavior effector class and different performance metrics (MSE, RMSE, CCC) have been used, hybrid decision-fusion appears to achieve the best prediction performances.

In general, different computational models tend to fit feature sets from different sources in unique ways. Decision-level fusion methods, as described in Section V.E, combine decisions from multiple computational models into a common decision, and their use should become more popular as number of sensor modalities within systems increases. Decision-level methods such as *set of rules* and *hybrid decision-fusion* have started to gain traction in conjunction with classification and regression models, respectively.

Although the number of features is a highly important factor in the training of computational models, nearly half of the reviewed works did not report this value. However, from those works that

did report it, the number of features ranges from 1 to ∼1000. On average, emotion recognition models tend to be trained with a higher number of features than models for the recognition of personality factors and social interactions, suggesting that emotion recognition systems are more computationally complex. Based on current studies, it is unclear if this computational complexity is linked to the complexity inherent to the personalization of human emotion. Emotion recognition models have also been more widely explored, and their complexity may be an artifact of the relative maturity of those models.

VII. Challenges and Opportunities

A. Theoretical Considerations

Human behavior monitoring is scenario-specific, requiring an understanding of how the behavior of interest manifests through nonverbal messages. Specific factors of human behavior need further study to better understand their effects during social interactions and their manifestations through nonverbal messages. As mentioned in Section II, we contend that there is a lack of research using behavior monitoring technologies to study the role of attitudes in human behavior during social interactions. This may result from a dominant paradigm existing in social sciences that treats humans as static actors rather than dynamic entities whose opinions change over time. We encourage future research to extend beyond this paradigm and explore how behavior monitoring technologies can be used to study the impact of attitudes. For example, attitudes towards other people in one's workgroup may influence how social interactions unfold, which could have an impact on the group's cohesiveness and decision-making effectiveness. Furthermore, positive and negative attitudes toward others in social interactions may affect the health and well-being of the participants.

In addition, we note that, technologies used for the monitoring of human emotions have been kept separate from those studying social interactions as well as personality factors. Because personality factors are directly related to trait affect and strongly tied to social interaction, we believe that future sensor monitoring systems should integrate aspects of existing emotion recognition systems and social wearable systems. Thus, we encourage further conversations with psychologists and cognitive engineers to advance this area of research.

B. Sensors, Pre-Processing, and Extraction of Signal Features

The development of wearable technologies is advancing rapidly. For example, functional near-infrared spectroscopy (fNIRS) [172]–[174] and time domain NIRS [175], [176] are effective brain imaging techniques could become as portable as EEGs. While such techniques are being miniaturized into wearable platforms, further exploration is needed to determine and validate their potential in providing valuable information for the real-time recognition of human behavior effectors.

Human behavior is greatly influenced by context and environment. However, the integration of environmental sensors and contextual information with behavior monitoring remains unexplored. On the other hand, we have seen that the integration of multiple sensor modalities involves the management of efficient pre-processing methods for sensor signals obtained at different timestamps, which is necessary because of variations in sampling frequency and windows of time traditionally used to keep computational complexity and memory consumption at its lowest. Thus, using multiple sensor modalities presents a challenge when integrating information from different sensor sources. First, the reviewed literature on physiological signals shows that pre-processing methods such as artifact removal have been largely ignored or manually applied. This reflects opportunities for research in sensor integration, wearability, and front-end processing for the reduction of noise and removal of signal artifacts. Second, windows of time used for emotion, personality factors, and social interaction prediction depend on the chosen sensor modality being used and on the behavior under observation. Although similarities in window length can be identified in the literature, there is still no clear consensus regarding the optimum window length for a given modality and behavior [177]. This remains an area of exploration to optimize human behavior monitoring systems.

Other challenges related to the integration of multiple sensor modalities include time-alignment of the collected multimodal sensor signals and variations of feature formats. For example, social interaction dynamics are composed of a mix of verbal and nonverbal messages that belong to different time scales and may not always be synchronized. We may convey a message using verbal communication and reinforce it with a gesture, however changes in acoustics will be captured more rapidly than gesture indicators because audio signals change more rapidly than physiological or movement signals. To deal with some of these challenges, methods such as data-level fusion, feature-level fusion, and decision-level fusion have been employed. Yet, optimized methods for employing these fusion mechanisms remains an open area of study for real-time human behavior monitoring systems.

Related to the extraction of features, it is valuable to know that, in some cases, to effectively reduce noise in data variability and reduce variation introduced by hardware, features are normalized. Data or feature normalization can ultimately improve overall human behavior recognition performance. This is particularly true when data from different hardware units and individuals are combined to select the right set of features. As noted above, selecting the most informative set of features has shown to improve overall system performance and computational complexity. However, further research is necessary to define efficient methods of human behavior data normalization for real-time processing.

C. Processing Methods

The heterogeneity of human behavior makes it challenging to establish a computational model that can fit everyone in all situations, and the fact that human behavior is highly dependent on context and environment further complicates the task. We observed that, overall, training subject-dependent models

(personalized models) tend to provide higher recognition accuracy than subject-independent models (generalize models) [23], [157], [160]. However, methods to generalize classifiers are of interest in order to reduce training time for a variety of real-time applications. To explore the design of generalized human behavior recognition models, computational models have been trained using training and testing datasets acquired from different individuals at different conditions [159]. Reinforcement learning techniques have also been applied to this goal of personalized human behavior recognition models [178], but this approach is still in the early stages of development and needs further exploration.

To date, most of the models that have been built ignore the facts that data changes over time, noise levels increase or decrease, and the source of the signals can also change (e.g., using different microphones). There have been efforts to tackle this problem by designing adaptive classification models, such as those presented in [159], but further advances are needed before such models could provide effective adaptation in real life applications. In addition, a thorough study of the limitations of different feature extraction methods and computational models, in terms of hardware resource allocation and suitability for real-time implementation, would greatly benefit the design of wearable human behavior monitoring systems.

D. Other Challenges in Human Behavior Monitoring Technologies

Direct comparison of the published work in human behavior monitoring is challenged by many inconsistencies in these studies, including different experimental setups for collection of training data, different number of subjects under study, and different human behaviors being elicited and studied. The automatic and real-time recognition of human behaviors is a complex problem. Because of this complexity, most training datasets have been collected under laboratory conditions, ignoring real-life elements that directly impact the human behavior being studied, and therefore system performance can degrade significantly when used in the field. This is particularly true for systems that intend to recognize emotions. Further work is needed in experimental design to develop training scenarios that lead to good performance when the systems are applied in real-life environments.

VIII. Conclusion

In this paper, we reviewed theoretical and technical aspects of human behavior monitoring technologies. Human behavior theory was discussed to illuminate the most relevant elements to the understanding of human behaviors of interest and how they can be monitored through physiological reactions and nonverbal messages. Based on a thorough review of the literature, all sensors used for the monitoring of human behaviors were identified and categorized, all features extracted from a variety of sensor signals were identified and summarized, and the features found to be highly correlated with specific behaviors were critically analyzed. In addition, the computational models used in behavior monitoring were reviewed, and those with the best performance were further analyzed while considering their applicability and challenges in real-time systems. Lastly, we have identified the challenges and opportunities for further research in real-time monitoring of human behaviors in real-life environments. This review shows that many systems, components, and methods for evaluating the behaviors and interaction of individuals have been explored, and, with continued effort, human behavior monitoring technologies will remain on the path to reach their full potential.

References

[1] S. N. Young, "The neurobiology of human social behaviour: An important but neglected topic," *J. Psychiatry Neurosci.*, vol. 33, no. 5, pp. 391–392, 2008.

[2] D. Umberson and J. K. Montez, "Social relationships and health: A flashpoint for health policy," *J. Health Soc. Behav.*, vol. 51, no. 1_suppl, pp. S54–S66, 2010.

[3] L. M. Hernandez and D. G. Blaze, Eds., "The impact of social and cultural environment on health," in *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate*, no. 2, National Academies Press (US), 2006.

[4] B. Uzzi, S. Mukherjee, M. Stringer, and B. Jones, "A typical combinations and scientific impact," *Science*, vol. 342, no. 6157, pp. 468–472, 2013.

[5] S. W. J. Kozlowski and D. R. Ilgen, "Enhancing the effectiveness of work groups and teams," *Psychol. Sci. Public Interest.*, vol. 7, no. 3, pp. 77–124, 2006.

[6] L. Lu, Y. C. Yuan, and P. L. McLeod, "Twenty-five years of hidden profiles in group decision making: A meta-analysis," *Pers. Soc. Psychol. Rev.*, vol. 16, no. 1, pp. 54–75, 2012.

[7] V. Rousseau, C. Aubé, and A. Savoie, "Teamwork behaviors: A review and an integration of frameworks," *Small Gr. Res.*, vol. 37, no. 5, pp. 540–570, 2006.

[8] K. Mastroianni and J. Storberg-Walker, "Do work relationships matter? Characteristics of workplace interactions that enhance or detract from employee perceptions of well-being and health behaviors," *Heal. Psychol. Behav. Med.*, vol. 2, no. 1, pp. 798–819, 2014.

[9] J. T. Cacioppo, L. C. Hawkley, G. J. Norman, and G. G. Berntson, "Social isolation," *Ann. N. Y. Acad. Sci.*, vol. 1231, no. 1, pp. 17–22, 2011.

[10] R. Mushtaq, S. Shoib, T. Shah, and S. Mushtaq, "Relationship between loneliness, psychiatric disorders and physical health ? A review on the psychological aspects of loneliness," *J. Clin. Diagn. Res.*, vol. 8, no. 9, pp. WE01–WE04, 2014.

[11] J. Holt-Lunstad, T. B. Smith, M. Baker, T. Harris, and D. Stephenson, "Loneliness and social isolation as risk factors for mortality: A meta-analytic review," *Perspect. Psychol. Sci.*, vol. 10, no. 2, pp. 227–237, 2015.

[12] Y. Hao, D. Wang, and J. G. Budd, "Design of intelligent emotion feedback to assist users regulate emotions: Framework and principles," in *Proc. Int. Conf. Affect. Comput. Intell. Interaction*, 2015, pp. 938–943.

[13] G. Chanel, S. Pelli, N. Ravaja, and K. Kuikkaniemi, "Social interaction using mobile devices and biofeedback: Effects on presence, attraction and emotions," in *Proc. BioSPlay Workshop, Fun Games Conf.*, 2010, pp. 5–9.

[14] J. Sturm, O. H. Herwijnen, A. Eyck, and J. Terken, "Influencing social dynamics in meetings through a peripheral display," in *Proc. 9th Int. Conf. Multimodal Interfaces*, 2007, pp. 263–270.

[15] T. Kim, D. O. Olguín, B. N. Waber, and A. Pentland, "Sensor-based feedback systems in organizational computing," in *Proc. Int. Conf. Comput. Sci. Eng.*, 2009, pp. 966–969.

[16] N. Alduncin, L. C. Huffman, H. M. Feldman, and I. M. Loe, "Executive function is associated with social competence in preschool-aged children born preterm or full term," *Early Hum. Dev.*, vol. 90, no. 6, pp. 299–306, 2014.

[17] R. Wheeler, "We all do it: Unconscious behavior, bias, and diversity," *Law Libr. J.*, vol. 107, no. 2, pp. 15–36, 2015.

[18] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.

[19] B. Buvaneswari and T. K. Reddy, "A review of EEG based human facial expression recognition systems in cognitive sciences," in *Proc. Int. Conf. Energy, Commun., Data Anal. Soft Comput.*, 2017, pp. 462–468.

[20] D. Gatica-Perez, "Analyzing group interactions in conversations: A review," in *Proc. IEEE Int. Conf. Multisensor Fusion Integr. Intell. Syst.*, 2006, pp. 41–46.

[21] X. Xu *et al.*, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 2074, pp. 1–41, 2018.

[22] M. Ali, A. H. Mosa, F. Al Machot, and K. Kyamakya, "Emotion recognition involving physiological and speech signals: A comprehensive review," in *Recent Advances in Nonlinear Dynamics and Synchronization*, K. Kyamakya, W. Mathis, R. Stoop, J. Chedjou, and Z. Li, Eds. Springer, Cham, 2018.

[23] S. Jing, X. Mao, and L. Chen, "Prominence features: Effective emotional features for speech emotion recognition," *Digit. Signal Process. A Rev. J.*, vol. 72, pp. 216–231, 2018.

[24] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A social signal processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[25] N. Lehmann-Willenbrock, H. Hung, and J. Keyton, "New frontiers in analyzing dynamic group interactions: Bridging social and computer science," *Small Gr. Res.*, vol. 48, no. 5, pp. 519–531, 2017.

[26] G. Chanel and C. Mühl, "Connecting brains and bodies: Applying physiological computing to support social interaction," *Interact. Comput.*, vol. 27, no. 5, pp. 534–550, 2015.

[27] M. I. S. Reddy, K. S. Reddy, V. U. Kumar, and P. V. V. Kumar, "Human computing and machine understanding of human behaviour: A survey," in *Artificial Intelligence for Human Computing*, T. S. Huang, A. Nijholt, M. Pantic, and A. Pentland, Eds. Springer, Berlin, Heidelberg, 2007, pp. 47–71.

[28] F. Cvrčková, V. Žárský, and A. Markoš, "Plant studies may lead us to rethink the concept of behavior," *Front. Psychol.*, vol. 7, pp. 10–13, 2016.

[29] A. Bandura, "Human agency in social cognitive theory," *Am. Psychol.*, vol. 44, no. 9, pp. 1175–1184, 1989.

[30] A. Bandura, "Social cognitive theory," in *Annals of child development*, vol. 6, R. Vasta, Ed. Greenwich, CT: JAI Press, pp. 1–60, 1989.

[31] W. Mischel, "The interaction of person and situation," in *Personality at the Crossroads: Current Issues in Interactional Psychology*, D. Magnusson and N. S. Endler, Eds. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc., 1977, pp. 333–352.

[32] T. J. Bouchard and J. C. Loehlin, "Genes, evolution, and personality," *Behav. Genet.*, vol. 31, no. 3, pp. 243–273, 2001.

[33] A. H. Eagly and S. Chaiken, *The Psychology of Attitudes*. Hardcourt Brace Jovanovich College Publishers, 1993.

[34] B. L. Fredrickson, "The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions," *Am. Psychol.*, vol. 56, no. 3, pp. 218–226, 2001.

[35] W. B. Cannon, "The James-Lange theory of emotions: A critical examination and an alternative theory," *Am. J. Psychol.*, vol. 39, no. 1, pp. 106–124, 1927.

[36] S. W. Porges, "Orienting in a defensive world: Mammalian modifications of our evolutionary heritage. A Polyvagal Theory," *Psychophysiology*, vol. 32, no. 4, pp. 301–318, 1995.

[37] S. W. Porges, "Polyvagal theory," *Biol. Psychol.*, vol. 74, no. 2, pp. 116–143, 2007.

[38] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[39] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.

[40] R. Plutchik, "The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *Am. Sci.*, vol. 89, no. 4, pp. 344–350, 2001.

[41] G. A. Van Kleef, "How emotions regulate social life: The Emotions as Social Information (EASI) model," *Curr. Dir. Psychol. Sci.*, vol. 18, no. 3, pp. 184–188, 2009.

[42] N. H. Frijda, *The Emotions*. Cambridge, MA, USA: Cambridge Univ. Press, 1986.

[43] P. L. Perrewé and P. E. Spector, "Personality research in the organizational sciences," *Res. Pers. Hum. Resour. Manag.*, vol. 21, G. R. Ferris and J. J. Martocchio, Eds. Elsevier Science/JAI Press, 2002, pp. 1–63.

[44] T. A. Judge, J. E. Bono, R. Ilies, and M. W. Gerhardt, "Personality and leadership: A qualitative and quantitative review," *J. Appl. Psychol.*, vol. 87, no. 4, pp. 765–780, 2002.

[45] M. C. Ashton and K. Lee, "Honesty-humility, the big five, and the five-factor model," *J. Pers.*, vol. 73, no. 5, pp. 1321–1354, 2005.

[46] M. C. Ashton and K. Lee, "Empirical, theoretical, and practical advantages of the HEXACO model of personality structure," *Pers. Soc. Psychol. Rev.*, vol. 11, no. 2, pp. 150–166, May 2007.

[47] D. L. Paulhus and K. M. Williams, "The dark triad of personality: Narcissism, Machiavellianism, and psychopathy," *J. Res. Pers.*, vol. 36, no. 6, pp. 556–563, Dec. 2002.

[48] J. B. Rotter, "Generalized expectancies for internal versus external control of reinforcement," *Psychol. Monogr. Gen. Appl.*, vol. 80, no. 1, pp. 1–28, 1966.

[49] J. K. Alberts, T. K. Nakayama, and J. N. Martin, *Human Communication in Society*, 3rd ed., Upper Saddle River, NJ, USA: Pearson, 2012.

[50] P. Watzlawick, J. B. Bavelas, and D. D. Jackson, *Pragmatics of Human Communication: A Study of Interactional Patterns, Pathologies and Paradoxes*. W. W. Norton Company, 1967.

[51] A. Pentland, *Honest Signals: How They Shape our World*. Cambridge, MA, USA: MIT Press, 2010.

[52] E. Goffman, *The Presentation of Self in Everyday Life*. New York, NY, USA: Anchor Books, 1959.

[53] C. Crivelli and A. J. Fridlund, "Facial displays are tools for social influence," *Trends Cogn. Sci.*, vol. 22, no. 5, pp. 388–399, 2018.

[54] A. Vinciarelli *et al.*, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 69–87, Jan.-Mar. 2012.

[55] A. Vinciarelli, H. Salamin, and M. Pantic, "Social signal processing: Understanding social interactions through nonverbal behavior analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, 2010, pp. 42–49.

[56] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans. Affect. Comput.*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[57] A. Pentland, "Social signal processing," *IEEE Signal Process. Mag.*, vol. 24, no. 4, pp. 108–111, 2007.

[58] G. Bente, "New tools – new insights: Using emergent technologies in nonverbal communication research," in *Reflections on Interpersonal Communication*, S. W. Wilson and S. W. Smith, Eds. San Diego, CA, USA: Cognella, 2019, pp. 161–188.

[59] K. Yun, K. Watanabe, and S. Shimojo, "Interpersonal body and neural synchronization as a marker of implicit social interaction," *Sci. Rep.*, vol. 2, no. 959, pp. 1–8, 2012.

[60] S. M. Thurman and H. Lu, "Perception of social interactions for spatially scrambled biological motion," *PLoS One*, vol. 9, no. 11, pp. 1–12, 2014.

[61] A. Innocenti, E. de Stefani, N. F. Bernardi, G. C. Campione, and M. Gentilucci, "Gaze direction and request gesture in social interactions," *PLoS One*, vol. 7, no. 5, pp. 1–8, 2012.

[62] E. De Stefani and D. De Marco, "Language, gesture, and emotional communication: An embodied view of social interaction," *Front. Psychol.*, vol. 10, no. 2063, pp. 1–8, 2019.

[63] C. Frith, "Role of facial expressions in social interactions," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1535, pp. 3453–3458, 2009.

[64] R. E. Jack and P. G. Schyns, "The human face as a dynamic tool for social communication," *Curr. Biol.*, vol. 25, no. 14, pp. R621–R634, 2015.

[65] P. Filippi, "Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language," *Front. Psychol.*, vol. 7, no. 1393, pp. 1–19, 2016.

[66] N. Henriksen, "Style, prosodic variation, and the social meaning of intonation," *J. Int. Phon. Assoc.*, vol. 43, no. 2, pp. 153–193, 2013.

[67] G. Bente, N. C. Kramer, and F. Eschenburg, "Is there anybody out there," *Mediat. Interpers. Commun.*, pp. 131–157, 2008.

[68] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: categories, origins, usage, and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[69] S. Duncan, "Some signals and rules for taking speaking turns in conversations," *J. Pers. Soc. Psychol.*, vol. 23, no. 2, pp. 283–292, 1972.

[70] G. Bente, H. Leuschner, A. Al Issa, and J. J. Blascovich, "The others: Universals and cultural specificities in the perception of status and dominance from nonverbal behavior," *Conscious. Cogn.*, vol. 19, no. 3, pp. 762–777, 2010.

[71] B. M. Depaulo and H. S. Friedman, "Nonverbal communication," in *The Handbook of Social Psychology*, D. T. Gilbert, S. T. Fiske, and G. Lindzey, Eds. New York, NY, USA: McGraw-Hill, 1998, pp. 3–40.

[72] A. Vinciarelli and A. S. Pentland, "New social signals in a new interaction world: The next frontier for social signal processing," *IEEE Syst. Man, Cybern. Mag.*, vol. 1, no. 2, pp. 10–17, Apr. 2015.

[73] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, no. 5, pp. 1203–1233, 2013.

[74] L. E. Holmquist, J. Falk, and J. Wigström, "Supporting group collaboration with interpersonal awareness devices," *Pers. Ubiquitous Comput.*, vol. 3, no. 1–2, pp. 13–21, 1999.

[75] R. Want, A. Hopper, V. Falcão, and J. Gibbons, "The active badge location system," *ACM Trans. Inf. Syst.*, vol. 10, no. 1, pp. 91–102, 1992.

[76] H. Jang, S. P. Choe, S. N. B. Gunkel, S. Kang, and J. Song, "A system to analyze group socializing behaviors in social parties," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 6, pp. 801–813, Dec. 2017.

[77] R. W. Devaul, S. J. Schwartz, and A. S. Pentland, "MIThril: Context-aware computing for daily life," *Web*, 2001, 2001. [Online]. Available: http://www.media.mit.edu/wearables/mithril/MIThril.pdf

[78] R. DeVaul, M. Sung, J. Gips, and A. Pentland, "MIThril 2003: Applications and architecture," in *Proc. 7th IEEE Int. Symp. Wearable Comput.*, 2003, pp. 4–11.

[79] N. Eagle and A. Pentland, "Wearables in the workplace: Sensing interactions at the office," in *Proc. IEEE Int. Symp. Wearable Comput.*, 2003, pp. 256–257.

[80] M. Laibowitz and J. A. Paradiso, "The UbER-Badge, a versatile platform at the juncture between wearable and social computing," in *Proc. Int. Conf. Pervasive Comput.*, 2004, pp. 1–6.

[81] J. Gips and A. Pentland, "Mapping human networks," in *Proc. 4th Annu. IEEE Int. Conf. Pervasive Comput. Commun.,* 2006, pp. 159–168.

[82] M. Laibowitz, J. Gips, R. Aylward, and A. Pentland, "A sensor network for social dynamics," in *Proc. 5th Int. Conf. Inf. Process. Sensor Netw.*, 2006, pp. 483–491.

[83] T. Choudhury and A. Pentland, "Characterizing social networks using the sociometer," in *Proc. North American Assoc. Comput. Social Org. Sci. (NAACSOS)*, 2004, pp. 1–4.

[84] W. Dong, B. Lepri, T. Kim, F. Pianesi, and A. S. Pentland, "Modeling conversational dynamics and performance in a social dilemma task," in *Proc. 5th Int. Symp. Commun. Control Signal Process.*, 2012, pp. 1–4.

[85] Y. Zhang *et al.*, "TeamSense: Assessing personal affect and group cohesion in small teams through dyadic interaction and behavior analysis with wearable sensors," in *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, 2018, pp. 1–22.

[86] T. Kim, A. Chang, L. Holland, and A. S. Pentland, "Meeting Mediator: Enhancing group collaboration using sociometric feedback," in *Proc. ACM Conf. Comput. Supported Cooperative Work*, 2008, pp. 457–466.

[87] O. Lederman, D. Calacci, A. MacMullen, D. C. Fehder, F. Murray, and A. S. Pentland, "Open Badges: A low-cost toolkit for measuring team communication and dynamics," pp. 1–7, 2017, *arXiv:1710.01842*.

[88] O. Lederman, A. Mohan, D. Calacci, and A. S. Pentland, "Rhythm: A unified measurement platform for human organizations," *IEEE Multimed.*, vol. 25, no. 1, pp. 26–38, Jan.-Mar. 2018.

[89] A. Madan and A. (Sandy) Pentland, "VibeFones: Socially aware mobile phones," in *Proc. 10th IEEE Int. Symp. Wearable Comput.*, 2006, pp. 109–112.

[90] J. Müller, S. Fàbregues, E. A. Guenther, and M. J. Romano, "Using sensors in organizational research-clarifying rationales and validation challenges for mixed methods," *Front. Psychol.*, vol. 10, no. 1188, pp. 1–14, 2019.

[91] J. Gu *et al.*, "Wearable social sensing: Content-based processing methodology and implementation," *IEEE Sens. J.*, vol. 17, no. 21, pp. 7167–7176, Nov. 2017.

[92] J. Gu *et al.*, "Wearable social sensing and its application in anxiety assesment," in *Proc. Int. Conf. Cyber-Enabled Distrib. Comput. Knowl. Discov.*, 2017, pp. 305–308.

[93] L. Jiang *et al.*, "Wearable long-term social sensing for mental wellbeing," *IEEE Sens. J.*, vol. 19, no. 19, pp. 8532–8542, Oct. 2019.

[94] J. Frey, M. Grabli, R. Slyper, and J. Cauchard, "Breeze: Sharing biofeedback through wearable technologies," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–12.

[95] L. Fraiwan, T. Basmaji, and O. Hassanin, "A mobile mental health monitoring system: A smart glove," in *Proc. - 14th Int. Conf. Signal Image Technol. Internet Based Syst.*, Jul. 2018, pp. 235–240.

[96] D. Girardi, F. Lanubile, and N. Novielli, "Emotion detection using noninvasive low cost sensors," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interaction*, 2017, pp. 125–130.

[97] R. S. McGinnis *et al.*, "Wearable sensors and machine learning diagnose anxiety and depression in young children," in *Proc. IEEE EMBS Int. Conf. Biomed. Health Inform.*, 2018, pp. 410–413.

[98] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," *Biocybern. Biomed. Eng.*, vol. 39, no. 2, pp. 444–469, 2019.

[99] N. Kehtarnavaz and M. Gamadia, "Real-time image and video processing: From research to reality," in *Proc. Synth. Lectures Image, Video, Multimedia Process.*, 2005, pp. 1–108.

[100] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affect. Comput.*, vol. 10, no. 3, pp. 325–347, Jul.-Sep. 2019.

[101] S. Yang *et al.*, "IoT structured long-term wearable social sensing for mental wellbeing," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 3652–3662, Apr. 2019.

[102] S. Ha *et al.*, "Integrated circuits and electrode interfaces for noninvasive physiological monitoring," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 5, pp. 1522–1537, May 2014.

[103] Y. Chuo *et al.*, "Mechanically flexible wireless multisensor platform for human physical activity and vitals monitoring," *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 5, pp. 281–294, Oct. 2010.

[104] R. Sharma, V. I. Pavlovic, and T. S. Huang, "Toward multimodal human-computer interface," *Proc. IEEE*, vol. 86, no. 5, pp. 853–869, 1998.

[105] A. Jaimes and N. Sebe, "Multimodal human-computer interaction: A survey," *Comput. Vis. Image Underst.*, vol. 108, no. 1–2, pp. 116–134, 2007.

[106] D. Gatica-Perez, *Modelling Interest in Face-to-Face Conversations From Multimodal Nonverbal Behaviour*, 1st ed. New York, NY, USA: Elsevier, 2010.

[107] M. Pantic, A. Nijholt, A. Pentland, and T. S. Huanag, "Human-centred intelligent human computer interaction (HCI^2): How far are we from attaining it?" *Int. J. Auton. Adapt. Commun. Syst.*, vol. 1, no. 2, p. 168, pp. 168–187, 2008.

[108] T. Theodorou, I. Mporas, and N. Fakotakis, "An overview of automatic audio segmentation," *Int. J. Inf. Technol. Comput. Sci.*, vol. 6, no. 11, pp. 1–9, Oct. 2014.

[109] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. 941–944.

[110] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[111] C. H. Wu and W. Bin Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Trans. Affect. Comput.*, vol. 2, no. 1, pp. 10–21, Jan. 2011.

[112] H. Balti and A. S. Elmaghraby, "Speech emotion detection using time dependent self organizing maps," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, 2013, pp. 470–478.

[113] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Proc. Speech Prosody*, 2012, pp. 1–4.

[114] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. Int. Conf. Spoken Lang. Process.*, 2002, pp. 2037–2040.

[115] S. Sahoo and A. Routray, "Detecting aggression in voice using inverse filtered speech features," *IEEE Trans. Affect. Comput.*, vol. 9, no. 2, pp. 217–226, Apr. 2018.

[116] D. B. Jayagopi and D. Gatica-Perez, "Mining group nonverbal conversational patterns using probabilistic topic models," *IEEE Trans. Multimed.*, vol. 12, no. 8, pp. 790–802, Dec. 2010.

[117] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Intell. Res.*, vol. 30, pp. 457–500, 2007.

[118] D. Furnham, "Language and Personality," in *Handbook of Language and Social Psychology*, H. Giles and W. Robinson, Eds. Chichester, U.K.: Winley, 1990.

[119] A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia, "From nonverbal cues to perception: Personality and social attractiveness," in *Proc. Cogn. Behav. Syst.*, 2012, pp. 60–72.

[120] K. Tusing, "The sounds of dominance. Vocal precursors of perceived dominance during interpersonal influence," *Hum. Commun. Res.*, vol. 26, no. 1, pp. 148–171, 2000.

[121] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 847–860, May 2011.

[122] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 17, no. 3, pp. 501–513, Mar. 2009.

[123] D. Sanchez-Cortes, O. Aran, M. S. Mast, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimed.*, vol. 14, no. 3, pp. 816–832, Jun. 2012.

[124] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: Training with unlabeled data," in *Proc. Companion Vol. Proc. HLT-NAACL - Short Papers*, 2003, pp. 34–36.

[125] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting group interest-level in meetings," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, pp. I489–I492.

[126] L. S. Kennedy and D. P. W. Ellis, "Pitch-based emphasis detection for characterization of meeting recordings," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2003, pp. 243–248.

[127] A. Cerekovic, O. Aran, and D. Gatica-Perez, "Rapport with virtual agents: What do human social cues and personality explain?" *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 382–395, Jul. 2017.

[128] D. B. Jayagopi, B. Raducanu, and D. Gatica-Perez, "Characterizing conversational group dynamics using nonverbal behaviour," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2009, pp. 370–373.

[129] A. Vinciarelli, "Capturing order in social interactions," *IEEE Signal Process. Mag.*, vol. 26, no. 5, pp. 133–152, Sep. 2009.

[130] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Trans. Multimed.*, vol. 12, no. 6, pp. 563–575, Oct. 2010.

[131] H. D. Critchley, "Electrodermal responses: What happens in the brain," *Neuroscientist*, vol. 8, no. 2, pp. 132–142, 2002.

[132] I. R. Kleckner *et al.*, "Simple, transparent, and flexible automated quality assessment procedures for ambulatory electrodermal activity data," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 7, pp. 1460–1467, Jul. 2018.

[133] S. Davila-Montero, S. Parsnejad, and A. J. Mason, "Exploring the relationship between speech and skin conductance for real-time arousal monitoring," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2020, pp. 1–5.

[134] P. Slovák, P. Tennent, S. Reeves, and G. Fitzpatrick, "Exploring skin conductance synchronisation in everyday interactions," in *Proc. 8th Nordic Conf. Hum.-Comput. Interaction: Fun, Fast, Found.*, 2014, pp. 511–520.

[135] H. J. Pijeira-Díaz, H. Drachsler, S. Järvelä, and P. A. Kirschner, "Investigating collaborative learning success with physiological coupling indices based on electrodermal activity," in *Proc. 6th Int. Conf. Learn. Anal. Knowl.*, 2016, pp. 64–73.

[136] E. Haataja, J. Malmberg, and S. Järvelä, "Monitoring in collaborative learning: Co-occurrence of observed behavior and physiological synchrony explored," *Comput. Hum. Behav.*, vol. 87, pp. 337–347, 2018.

[137] H. J. Pijeira-Díaz, H. Drachsler, S. Järvelä, and P. A. Kirschner, "Sympathetic arousal commonalities and arousal contagion during collaborative learning: How attuned are triad members?" *Comput. Hum. Behav.*, vol. 92, pp. 188–197, 2019.

[138] "10/20 System Positioning Manual," Accessed: Jun. 30, 2020, 2012. [Online]. Available: https://www.trans-cranial.com/docs/10_20_pos_man_v1_0_pdf.pdf

[139] V. Jurcak, D. Tsuzuki, and I. Dan, "10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems," *Neuroimage*, vol. 34, no. 4, pp. 1600–1611, Feb. 2007.

[140] M. Teplan, "Fundamentals of EEG measurement," *Meas. Sci. Rev.*, vol. 2, no. 2, pp. 1–11, 2002.

[141] S. Valenzi, T. Islam, P. Jurica, and A. Cichocki, "Individual classification of emotions using EEG," *J. Biomed. Sci. Eng.*, vol. 07, no. 08, pp. 604–620, 2014.

[142] L. Zou, X. Chen, G. Dang, Y. Guo, and Z. J. Wang, "Removing muscle artifacts from EEG data via underdetermined joint blind source separation: A simulation study," *IEEE Trans. Circuits Syst. II Exp. Briefs*, vol. 67, no. 1, pp. 187–191, Jan. 2020.

[143] Y. Yang and J. Zhou, "Recognition and analyses of EEG&ERP signals related to emotion: From the perspective of psychology," in *Proc. First Int. Conf. Neural Interface Control*, 2005, pp. 96–99.

[144] R. N. Duan, J. Y. Zhu, and B. L. Lu, "Differential entropy feature for EEG-based emotion classification," in *Proc. Int. IEEE/EMBS Conf. Neural Eng.*, 2013, pp. 81–84.

[145] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 327–339, Jul.-Sep. 2014.

[146] W. L. Zheng and B. L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Trans. Auton. Ment. Dev.*, vol. 7, no. 3, pp. 162–175, Sep. 2015.

[147] B. Pholpoke, T. Songthawornpong, and W. Wattanapanitch, "A micropower motion artifact estimator for input dynamic range reduction in wearable ECG acquisition systems," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 5, pp. 1021–1035, Oct. 2019.

[148] U. Satija, B. Ramkumar, and M. Sabarimalai Manikandan, "A review of signal processing techniques for electrocardiogram signal quality assessment," *IEEE Rev. Biomed. Eng.*, vol. 11, pp. 36–52, 2018.

[149] D. Nikolova, P. Petkova, A. Manolova, and P. Georgieva, "ECG-based emotion recognition: Overview of methods and applications," in *Proc. ANNA '18; Adv. Neural Netw. Appl.*, 2018, pp. 1–5.

[150] C. Xiefeng, Y. Wang, S. Dai, P. Zhao, and Q. Liu, "Heart sound signals can be used for emotion recognition," *Sci. Rep.*, vol. 9, no. 1, Dec. 2019.

[151] J. Cai, G. Liu, and M. Hao, "The research on emotion recognition from ECG signal," in *Proc. Int. Conf. Inf. Technol. Comput. Sci.*, 2009, pp. 497–500.

[152] D. S. Quintana, A. J. Guastella, T. Outhred, I. B. Hickie, and A. H. Kemp, "Heart rate variability is associated with emotion recognition: Direct evidence for a relationship between the autonomic nervous system and social cognition," *Int. J. Psychophysiol.*, vol. 86, no. 2, pp. 168–172, 2012.

[153] R. Gravina, P. Alinia, H. Ghasemzadeh, and G. Fortino, "Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges," *Inf. Fusion*, vol. 35, pp. 1339–1351, 2017.

[154] J. Kim, "Bimodal emotion recognition using speech and physiological changes," in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. 2007, Vienna, Austria: I-Tech, pp. 265–280.

[155] G. Chanel, M. Bétrancourt, T. Pun, D. Cereghetti, and G. Molinari, "Assessment of computer-supported collaborative processes using interpersonal physiological and eye-movement coupling," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interaction*, 2013, pp. 116–122.

[156] G. Chanel, S. Avry, G. Molinari, M. Betrancourt, and T. Pun, "Multiple users' emotion recognition: Improving performance by joint modeling of affective reactions," in *Proc. 7th Int. Conf. Affect. Comput. Intell. Interaction*, 2017, pp. 92–97.

[157] T. Vogt, E. André, and N. Bee, "EmoVoice — A framework for online recognition of emotions from voice," in *Perception in Multimodal Dialogue Systems. PIT 2008. Lecture Notes in Computer Science*, vol. 5078, E. André, L. Dybkjær, W. Minker, H. Neumann, R. Pieraccini, and M. Weber, Eds. Berlin, Heidelberg: Springer, 2008.

[158] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan.-Mar. 2012.

[159] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2015, pp. 5058–5062.

[160] L. Cen, F. Wu, Z. L. Yu, and F. Hu, *A Real-Time Speech Emotion Recognition System and its Application in Online Learning*. New York, NY, USA: Elsevier Inc., 2016.

[161] R. Rajak and R. Mall, "Emotion recognition from audio, dimensional and discrete categorization using CNNs," in *Proc. IEEE Region 10th Conf. TENCON*, 2019, pp. 301–305.

[162] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques," *Procedia Comput. Sci.*, vol. 49, pp. 50–57, 2015.

[163] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 273–284, Jul.-Sep. 2012.

[164] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–317, Mar. 2005.

[165] K. Katevas, K. Hänsel, R. Clegg, I. Leontiadis, H. Haddadi, and L. Tokarchuk, "Finding dory in the crowd: Detecting social interactions using multi-modal mobile sensing," in *Proc. 1st Workshop Mach. Learn. Edge Sensor Syst.*, 2019, pp. 37–42.

[166] M. Wöllmer *et al.*, "Abandoning emotion classes - Towards continuous emotion recognition with modelling of long-range dependencies," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 597–600.

[167] F. Ringeval *et al.*, "The AV + EC 2015: The first affect recognition challenge bridging across audio, video, and physiological data," in *Proc. 5th Int. Workshop Audio/Vis. Emotion Challenge*, 2015, pp. 3–8.

[168] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[169] K. Brady *et al.*, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, 2016, pp. 97–104.

[170] W. L. Zheng and B. L. Lu, "A multimodal approach to estimating vigilance using EEG and forehead EOG," *J. Neural Eng.*, vol. 14, no. 2, pp. 1–14, 2017.

[171] A. Huk, K. Bonnen, and B. J. He, "Beyond trial-based paradigms: Continuous behavior, ongoing neural activity, and natural stimuli," *J. Neurosci.*, vol. 38, no. 35, pp. 7551–7558, Aug. 2018.

[172] M. J. Saikia, W. G. Besio, and K. Mankodiya, "WearLight: Toward a wearable, configurable functional NIR spectroscopy system for noninvasive neuroimaging," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 1, pp. 91–102, Feb. 2019.

[173] M. A. Yaqub, S. W. Woo, and K. S. Hong, "Compact, portable, high-density functional near-infrared spectroscopy system for brain imaging," *IEEE Access*, vol. 8, pp. 128224–128238, 2020.

[174] J. Xu *et al.*, "A 665 μw silicon photomultiplier-based NIRS/EEG/EIT monitoring ASIC for wearable functional brain imaging," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 6, pp. 1267–1277, Dec. 2018.

[175] E. Conca *et al.*, "Large-area, fast-gated digital SiPM with integrated TDC for portable and wearable time-domain NIRS," *IEEE J. Solid-State Circuits*, vol. 55, no. 11, pp. 3097–3111, Nov. 2020.

[176] S. Saha, Y. Lu, F. Lesage, and M. Sawan, "Wearable SiPM-based NIRS interface integrated with pulsed laser source," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1313–1323, Dec. 2019.

[177] F. Ringeval *et al.*, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognit. Lett.*, vol. 66, pp. 22–30, 2015.

[178] O. O. Rudovic, M. Zhang, B. Schuller, and R. W. Picard, "Multi-modal active learning from human data: A deep reinforcement learning approach," in *Proc. Int. Conf. Multimodal Interaction*, 2019, pp. 6–15.

[179] F. Moukayed, H. Yun, T. Bisson, and A. Fortenbacher, "Detecting academic emotions from learners' skin conductance and heart rate: A data-driven approach using fuzzy logic," in *Proc. DeLFI Workshops*, 2018, pp. 1–10.

[180] B. Zhong *et al.*, "Emotion recognition with facial expressions and physiological signals," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2017, pp. 1–8.

[181] H. Yun, A. Fortenbacher, N. Pinkwart, T. Bisson, and F. Moukayed, "A pilot study of emotion detection using sensors in a learning context: Towards an affective learning companion," in *DeLFI/GMW Workshops*, 2017, pp. 1–11.

[182] L. Zhang *et al.*, "'BioVid Emo DB': A multimodal database for emotion analyses validated by subjective ratings," in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2017, pp. 1–6.

[183] W. Mou, H. Gunes, and I. Patras, "Alone versus In-a-group : A comparative analysis of facial affect recognition," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 521–525.

[184] W. Wei and Q. Jia, "Weighted feature gaussian kernel SVM for emotion recognition," *Comput. Intell. Neurosci.*, vol. 2016, pp. 1–8, 2016.

[185] S. Chen, Y. L. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features," *Image Vis. Comput.*, vol. 31, no. 2, pp. 175–185, 2013.

[186] S. Koelstra *et al.*, "DEAP: A database for emotion analysis using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 18–31, Jan.-Mar. 2012.

[187] M. Shimura, F. Monma, S. Mitsuyoshi, M. Shuzo, T. Yamamoto, and I. Yamada, "Descriptive analysis of emotion and feeling in voice," in *Proc. 6th Int. Conf. Natural Lang. Process. Knowl. Eng.*, 2010, pp. 1–4.

[188] J.-C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, and S. Abrilian, "Manual annotation and automatic image processing of multimodal emotional behaviors: Validating the annotation of TV interviews," *Pers. Ubiquitous Comput.*, vol. 13, no. 1, pp. 69–76, 2009.

[189] M. M. Khan, R. D. Ward, and M. Ingleby, "Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature," *ACM Trans. Appl. Percept.*, vol. 6, no. 1, pp. 1–22, 2009.

[190] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2008, pp. 1–6.

[191] Y. S. Shin, "Facial expression recognition based on emotion dimensions on manifold learning," in *Proc. Int. Conf. Comput. Sci.*, 2007, pp. 81–88.

[192] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2005, pp. 5–8.

[193] L. Chaby, M. Chetouani, M. Plaza, and D. Cohen, "Exploring multimodal social-emotional behaviors in autism spectrum disorders: An interface between social signal processing and psychopathology," in *Proc. ASE/IEEE Int. Conf. Privacy, Secur., Risk Trust ASE/IEEE Int. Conf. Social Comput., SocialCom/PASSAT*, 2012, pp. 950–954.

[194] A. Mahdhaoui, F. Ringeval, and M. Chetouani, "Emotional speech characterization based on multi-features fusion for face-to-face interaction," in *Proc. Int. Conf. Signals, Circuits Syst.*, 2009, pp. 1–6.

[195] M. Dahmane, P.-L. St-Charles, M. Lalonde, K. Heffner, and S. Foucher, "Arousal and valence estimation for visual non-intrusive stress monitoring," in *Proc. 9th Int. Conf. Image Process. Theory, Tools Appl.*, 2019, pp. 1–6.

[196] F. Ahmed, A. S. M. H. Bari, and M. L. Gavrilova, "Emotion recognition from body movement," *IEEE Access*, vol. 8, pp. 11761–11781, 2020.

[197] B. D. Yetton, J. Revord, S. Margolis, S. Lyubomirsky, and A. R. Seitz, "Cognitive and physiological measures in well-being science: Limitations and lessons," *Front. Psychol.*, vol. 10, no. 1630, pp. 1–18, 2019.

[198] T. Keshari and S. Palaniswamy, "Emotion recognition using feature-level fusion of facial expressions and body gestures," in *Proc. 4th Int. Conf. Commun. Electron. Syst.*, 2019, pp. 1184–1189.

[199] M. Raja and S. Sigg, "RFexpress! - RF emotion recognition in the wild," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops, PerCom Workshops*, 2017, pp. 38–41.

[200] A. Pradhan, A. Singh, and S. Saraswat, "Emotion recognition through wireless signal," in *Proc. 4th Int. Conf. Signal Process. Integr. Netw.*, 2017, pp. 91–95.

[201] C. Beyan, F. Capozzi, C. Becchio, and V. Murino, "Prediction of the leadership style of an emergent leader using audio and visual nonverbal features," *IEEE Trans. Multimed.*, vol. 20, no. 2, pp. 441–456, Feb. 2018.

[202] L. Batrinca, N. Mana, B. Lepri, N. Sebe, and F. Pianesi, "Multimodal personality recognition in collaborative goal-oriented tasks," *IEEE Trans. Multimed.*, vol. 18, no. 4, pp. 659–673, Apr. 2016.

[203] O. Aran and D. Gatica-Perez, "Fusing audio-visual nonverbal cues to detect dominant people in small group conversations," in *Proc. Int. Conf. Pattern Recognit.*, 2010, pp. 3687–3690.

[204] F. Pianesi, B. Lepri, A. Cappelletti, M. Zancanaro, and N. Mana, "Multimodal recognition of personality traits in social interactions," in *Proc. 10th Int. Conf. Multimodal Interfaces*, 2008, pp. 53–60.

[205] H. Hung *et al.*, "Using audio and video features to classify the most dominant person in a group meeting," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 835–838.

[206] D. Zhang, D. Gatica-perez, S. Bengio, and D. Roy, "Learning influence among interacting Markov chains," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 1577–1584.

[207] S. Basu, T. Choudhury, B. Clarkson, and A. Pentland, "Towards measuring human interactions in conversational settings," in *Proc. IEEE Int. Workshop Cues Commun.*, 2001, pp. 1577–1584.

[208] Z. Shen, A. Elibol, and N. Y. Chong, "Inferring human personality traits in human-robot social interaction," in *Proc. 14th ACM/IEEE Int. Conf. Hum.-Robot Interaction*, 2019, pp. 578–579.

[209] G. Leone, S. Migliorisi, and I. Sessa, "Detecting social signals of honesty and fear of appearing deceitful: A methodological proposal," in *Proc. 7th IEEE Int. Conf. Cogn. Infocommun.*, 2016, pp. 289–294.

[210] L. Ahonen, B. U. Cowley, A. Hellas, and K. Puolamäki, "Biosignals reflect pair-dynamics in collaborative work: EDA and ECG study of pair-programming in a classroom environment," *Sci. Rep.*, vol. 8, no. 1, pp. 1–16, 2018.

[211] J. Malmberg, S. Järvelä, J. Holappa, E. Haataja, X. Huang, and A. Siipo, "Going beyond what is visible: What multichannel data can reveal about interaction in the context of collaborative learning?" *Comput. Hum. Behav.*, vol. 96, pp. 235–245, 2019.

[212] M. T. Knierim, D. Jung, V. Dorner, and C. Weinhardt, "Designing live biofeedback for groups to support emotion management in digital collaboration," in *Proc. Int. Conf. Des. Sci. Res. Inf. System Technol.*, 2017, pp. 479–484.
[213] A. Sandy Pentland, "Social dynamics: Signals and behavior," in *Proc. Int. Conf. Devlop. Learn.*, 2004, pp. 1–5.
[214] A. Marcos-Ramiro, D. Pizarro, M. Marron-Romera, and D. Gatica-Perez, "Let your body speak: Communicative cue extraction on natural interaction using RGBD data," *IEEE Trans. Multimed.*, vol. 17, no. 10, pp. 1721–1732, Oct. 2015.
[215] E. Shmueli, V. K. Singh, B. Lepri, and A. Pentland, "Sensing, understanding, and shaping social behavior," *IEEE Trans. Comput. Soc. Syst.*, vol. 1, no. 1, pp. 22–34, Mar. 2014.
[216] U. Avci and O. Aran, "Effect of nonverbal behavioral patterns on the performance of small groups," in *Proc. Workshop Understanding Model. Multiparty, Multimodal Interact.*, 2014, pp. 9–14.
[217] J. Terken, J. Sturm, and I. Patras, "Multimodal support for social dynamics in co-located meetings," *Pers. Ubiquitous Comput.*, vol. 14, no. 8, pp. 703–714, 2010.
[218] C. Busso, P. G. Georgiou, and S. S. Narayanan, "Real-time monitoring of participants' interaction in a meeting using audio-visual sensors," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2007, pp. 685–688.
[219] Y. Shi, S. Das, S. Douglas, and S. Biswas, "An experimental wearable IoT for data-driven management of autism," in *Proc. 9th Int. Conf. Commun. Syst. Netw.*, Jun. 2017, pp. 468–471.
[220] G. Schiavo, "Socially-aware interfaces for supporting collocated interaction," in *Proc. IUI Companion '14: Proc. Companion Publication 19th Int. Conf. Intell. User Interfaces*, 2014, pp. 65–67.
[221] S. O. Ba and J. M. Odobez, "Multiperson visual focus of attention from head pose and meeting contextual cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 1, pp. 101–116, Jan. 2011.
[222] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "Identifying human behaviors using synchronized audio-visual cCues," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 54–66, Jan. 2017.
[223] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *Proc. IEEE Int. Conf. Privacy, Secur., Risk, Trust, IEEE Int. Conf. Social Comput.*, 2011, pp. 613–616.
[224] M. T. Curran, J. R. Gordon, L. Lin, P. K. Sridhar, and J. Chuang, "Understanding digitally-mediated empathy: An exploration of visual, narrative, and biosensory informational cues," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.
[225] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vis. Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.
[226] S. W. Byun, S. P. Lee, and H. S. Han, "Feature selection and comparison for the emotion recognition according to music listening," in *Proc. Int. Conf. Robot. Automat. Sci.*, 2017, pp. 172–176.

Sylmarie Dávila-Montero (Student Member, IEEE) received the B.S. degree in electrical engineering (with high honors) from the University of Puerto Rico, Mayagüez, Puerto Rico, in May 2015. Since Fall 2015, she has been working toward the Ph.D. degree in electrical engineering with Michigan State University, East Lansing, MI, USA, working under the supervision of Dr. Andrew J. Mason. Her research interests include real-time processing of biomedical and social signals, efficient implementation of machine learning algorithms, wearable health sensors, human–machine interfaces, the design of efficient real-time feature extraction, and signal processing algorithms for smart social sensing platforms.

Jocelyn Alisa Dana-Lê received the B.S. degree in psychology and management and society from the University of North Carolina, Chapel Hill, NC, USA, in 2016. She is currently an Organizational Behavior Doctoral Candidate with the Management Department, Michigan State University, East Lansing, MI, USA. Her research and teaching interests include work meaningfulness, vocational behavior, and alternative work arrangements. Her current research examines emotions and well-being, work–family spillover, and telework. She is a Member of the Society for Industrial and Organizational Psychology, the Southern Management Association, and the Academy of Management.

Gary Bente studied psychology from the University of Regensburg, Regensburg, Germany, and Saarbrücken, Germany. He received the Ph.D. degree in psychology, after his training as a Clinical Psychologist, from the University of Trier, Trier, Germany, in 1985. From 1991 to 2017, he was a Professor of psychology with the University of Cologne, Cologne, Germany. Since 2017, he has been a Professor with the Department of Communication, Michigan State University, East Lansing, MI, USA, and the Director with the Center for Avatar Research and Immersive Social Media Applications. He has authored or coauthored more than 300 papers in peer-reviewed journals and numerous book chapters. He is the Co-Editor of the German textbook, Media Psychology, in which he covered the chapters on psychophysiological measurement and eye tracking. His interdisciplinary research interests include computer science, engineering and neuroscience focuses on nonverbal communication, and the use of virtual reality technologies to study the cognitive and emotional processes in social interaction. He was the Editor-in-Chief of the *Journal of Media Psychology* for many years. He is currently working on an NSF-funded research project, applying machine learning algorithms to motion capture, eye tracking, and neurophysiological data to identify behavioral and biological signatures of interpersonal synchrony and rapport.

Angela T. Hall received the J.D. degree in 1993 from the College of Law, Florida State University, Tallahassee, FL, USA, and the Ph.D. degree in business administration from the College of Business, Florida State University, Tallahassee, FL, USA, in 2005. She is currently an Associate Professor with the School of Human Resources and Labor Relations, Michigan State University (MSU), East Lansing, MI, USA. Prior joining MSU, she was on the faculties of the University of Texas at San Antonio, San Antonio, TX, USA, and Florida State University, Tallahassee, FL, USA. Her research has appeared in various journals, such as the *Journal of Organizational Behavior, the Human Relations, the Personnel Psychology, and the Organizational Behavior and Human Decision Processes*. Her research interests include employee accountability, employee legal claiming, social influence, organizational politics, impression management, technology at work, diversity and inclusion, and barriers to employment. She is currently an Associate Editor for the *Africa Journal of Management.*

Andrew J. Mason (Senior Member, IEEE) received the B.S. degree in physics with highest distinction from Western Kentucky University, Bowling Green, KY, USA, in 1991, the BSEE degree with honors from the Georgia Institute of Technology, Atlanta, GA,USA, in 1992, and the MS and Ph.D. degrees in electrical engineering from The University of Michigan, Ann Arbor, MI, USA, in 1994 and 2000, respectively. From 1999 to 2001, he was an Assistant Professor with the University of Kentucky, Lexington, KY, USA. In 2001, he joined the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA, where he is currently a Professor. His research explores technologies for augmented human awareness and biomedical applications, including microfabricated structures, mixed-signal and embedded circuits, and machine learning algorithms. His current projects are focused on the design of human augmentation technologies, such as wearable biochemical, environmental and social sensing platforms, signal processing algorithms, and hardware for brain–machine interface and rapid unobtrusive machine-to-human communication. He is with the Sensory Systems and the Biomedical Circuits and Systems Technical Committees of the IEEE Circuits and Systems Society. He is an Associate Editor for the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS and is regularly on the technical and review committees for several IEEE conferences. He was the Co-General Chair of the 2011 IEEE Biomedical Circuits and Systems Conference. He was the recipient of the 2006 Michigan State University Teacher-Scholar Award and the 2010 Withrow Award for Teaching Excellence.