CHAPTER 5

Automated Mental State Detection for Mental Health Care

Sidney K. D'Mello
Departments of Psychology and Computer Science, University of Notre Dame, Notre Dame, IN, USA

INTRODUCTION

Jackson, a 30-year-old male, was recently diagnosed with depression and is receiving weekly cognitive-behavioral therapy (CBT) sessions. He wears a device that records his physiological activity (e.g., electrodermal activity, heart rate variability) as he goes about his everyday routine. At the onset of each CBT session, his therapist inputs his physiological data for the week into a computer program. The program provides an aggregate of Jackson's levels of positive and negative affect as well as pinpoints moments where these responses peaked. The therapist uses this information to monitor Jackson's progress and to dynamically tailor the therapy, for example, by asking Jackson to recall events corresponding to some of the peaks in affect.

Olivia is a 3-year-old girl who recently dislocated her right shoulder. She feels some pain, but has difficulty precisely expressing the intensity of the felt pain to her physician. The physician asks her to hold still for 5 s and then to rotate her arm using standard range of motion tests (e.g., abduction, flexion). This makes her wince and she utters a small cry. The physician records a video of her face both when her arm was still and during the tests. An audio recording of her cry is also synchronized with the video. A computer program provides an estimate of Olivia's pain intensity by analyzing the change in her facial expressions and sound patterns during the range of motion tests compared to when her arm was held still. In addition to the X-ray taken earlier, and his physical exam, he uses the computer-provided estimate of Olivia's pain intensity to select a course of treatment.

Luis was recently diagnosed with attention deficit hyperactivity disorder (ADHD). He simply cannot concentrate on his schoolwork and his grades are beginning to suffer. This causes him to feel depressed

Artificial Intelligence in Behavioral and Mental Health Care.
DOI: http://dx.doi.org/10.1016/B978-0-12-420248-1.00005-2

© 2016 Elsevier Inc.
All rights reserved.

because he thinks he is not "smart enough" to succeed in college. Luis's psychiatrist prescribes him with a standard amphetamine stimulant. He also gives Luis a small device that can be affixed to his laptop. The device monitors Luis's eye gaze when he uses his laptop to complete homework assignments, do assigned readings, or to study for a test. A computer program automatically analyzes his eye gaze and provides estimates of Luis's levels of attention during various phases in the study session. For example, it estimates when Luis is concentrating, when he is distracted, and even when he zones out despite trying to focus. The software provides Luis with feedback on his levels of attention aligned with the various study activities. Luis uses this information to replan his study strategy and to restudy certain topics. In addition, a new software program he downloaded can actually use this information in real-time to suggest topics for restudy. Luis's grades start to improve and he feels empowered.

These hypothetical scenarios illustrate how machines (devices and computer programs) that can automatically detect a person's mental state can provide actionable information to improve mental health care. These machine-provided mental state estimates can complement self- or observer-reports of the same constructs as in the cases of Jackson and Olivia. They can also afford reflection and dynamic action as with Luis. These machines can detect a variety of affective and cognitive states, such as physiological arousal, feelings (e.g., pain), affective dimensions (e.g., valence and arousal), cognitive states (e.g., attention and mental workload), emotional states (e.g., sad, happy, angry), and even complex cognitive-affective blends (e.g., confusion, frustration). They can target both momentary episodes (e.g., specific attentional lapses), intermediate mood states (e.g., a bad day), and prolonged mental phenomena (e.g., stress and depression). Some are best suited for use in controlled settings (e.g., a physician's office), others are best suited for home and office use, while some can be deployed in the wild for long-term ambulatory monitoring of mental states.

The "AI" in automated mental state detection comes from the overall goal of developing machines with the capability of sensing complex mental phenomena, which was previously a uniquely human ability, and from the different subfields of AI involved in the development of such machines (e.g., computer vision, machine learning). The purpose of this chapter is to demystify these seemingly magical machines that can "read-out" a person's mental state. This is done by first providing the theoretical and technical foundation of the highly interdisciplinary field of "automatic mental state detection." This is followed by an exposition of a few illustrative examples of recent mental state detection systems. The chapter concludes with a discussion of open issues in the field and provides some speculative comments on its future.

THEORETICAL AND TECHNICAL FOUNDATION

Automated detection of mental states is an active area of research within the broader umbrella of *human–computer interaction* and its sister field of *human factors and cognitive ergonomics*. It is composed of different subfields, such as social signal processing (Mehu & Scherer, 2012; Vinciarelli, Pantic, & Bourlard, 2009), affective computing (Cowie et al., 2001; Douglas-Cowie et al., 2007; Marsella, Gratch, & Petta, 2010; McKeown, Valstar, Cowie, Pantic, & Schroder, 2012; Picard, 1997, 2010), attention-aware computing (D'Mello, Cobian, & Hunter, 2013; Roda & Thomas, 2006), and augmented cognition (Marshall, 2005; St. John, Kobus, Morrison, & Schmorrow, 2004), each focusing on different mental states in different contexts. Automated mental state detection is truly an interdisciplinary field. Its psychological roots are in cognitive psychology, affective sciences, social psychology, study of nonverbal behavior, and psychophysiology. Its technical roots lie in engineering and computer science, specifically sensors and wearable devices, digital signal processing (e.g., computer vision, acoustic modeling), and machine learning.

The psychological arm of automated mental state detection is grounded in theories that highlight the embodied nature of mental processes. Embodied theories of cognition and affect posit that mental states are not restricted to the confines of the mind but are manifested in the body (Barsalou, 2008; deVega, Glenberg, & Graesser, 2008; Ekman, 1992; Niedenthal, 2007; Russell, Bachorowski, & Fernandez-Dols, 2003). One of the most direct examples of a mind–body link is the increased activation of the sympathetic nervous system during fight-or-flight responses (Larsen, Berntson, Poehlmann, Ito, & Cacioppo, 2008). There are also well-known relationships between facial expressions and affective states (Coan, 2010; Ekman, 1984), for example, the furrowed brow during experiences of confusion (Darwin, 1872; D'Mello & Graesser, 2014). There is also a long history of using bodily/physiological responses to investigate cognitive processes like attention and cognitive load. For example, the study of eye movements (oculesics) has emerged as an invaluable tool to investigate visual attention (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Rayner, 1998), while

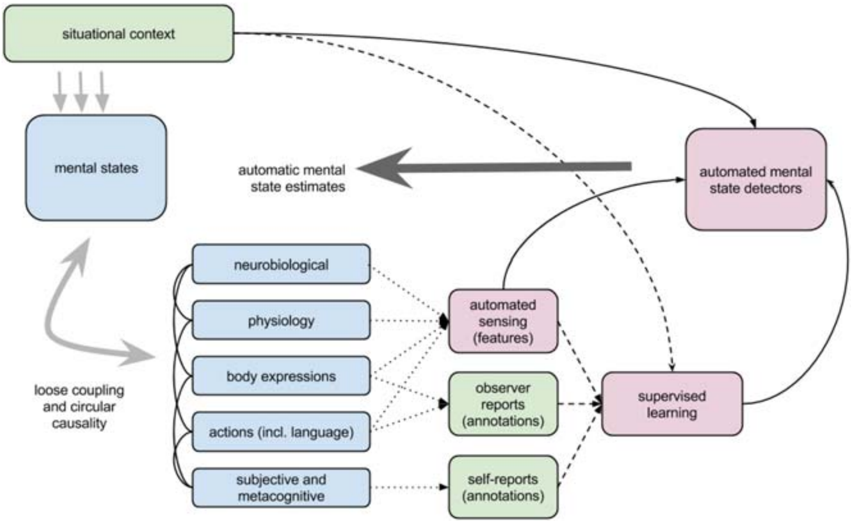
situational context
mental states
automatic mental state estimates
automated mental state detectors
neurobiological
physiology
body expressions
actions (incl. language)
subjective and metacognitive
automated sensing (features)
observer reports (annotations)
self-reports (annotations)
supervised learning
loose coupling and circular causality

Figure 5.1 Processes and steps in automatic mental state detection.

electroencephalography has long been used as an index into mental workload (Berka et al., 2007). This close mind–body relationship is unsurprising once one realizes that cognition and affect are in the service of action. Simply put, we think and we feel in order to act. Bodies are the agents of action, hence, monitoring observable bodily changes can provide critical insights into unobservable mental states. This key idea underlies the automated detection approach which attempts to infer mental states from bodily responses.

Figure 5.1 provides a summary of the foundational ideas of automated mental state detection. The general idea is that a person's interactions with the world (situational context) give rise to latent (or hidden) mental states that cannot be directly measured. The mental states are associated with changes at multiple levels (neurobiological, physiological, bodily expressions, overt actions, and subjective/metacognitive feelings/reflections), which in turn influence the mental states themselves — this is referred to as circular causality (Lewis, 2005). Some of these changes are implicit (e.g., neurobiological, some physiological changes) in that they occur outside of conscious awareness, while others are more explicit (e.g., overt actions, metacognitive reflections). A subset of these implicit and explicit changes is readable by machine sensors and human observers, while others are only accessible to the self (dotted lines in Figure 5.1).

The computational problem is to infer (or estimate) the latent mental states from the machine-readable signals recorded by the sensors.

Solving the aforementioned inference problem requires solving two interrelated computational challenges. The first challenge is to extract diagnostic information (called features) from the signals recorded by the sensors. This is a sensor-specific component of mental state detection because the method used varies based on the sensor and corresponding signal. For example, if the sensor is a webcam, then the signal is a video (presumably of the face). In this case, the features might be the activation of specific facial muscles or Action Units (AUs) (Ekman & Friesen, 1978), such as the inner brow raise (AU 1) or the lip pucker (AU 18). Computer vision-based techniques are needed to automatically compute these facial features from video (Pantic & Patras, 2006; Valstar, Mehu, Jiang, Pantic, & Scherer, 2012). Similarly, pitch and amplitude are common paralinguistic (acoustic-prosodic) features extracted from an audio signal recorded with a microphone (the sensor). Here, digital signal processing methods applied to the speech domain are needed (Eyben, Wöllmer, & Schuller, 2010). If the content of the spoken signal is to be analyzed, then this requires automatic speech recognition followed by natural language processing techniques to identify pertinent features (Pang & Lee, 2008).

The second challenge involves making inferences of a person's mental state from the features extracted from the signals. This falls under the purview of machine learning and is somewhat, but not entirely, signal-independent. Most (but not all) researchers adopt supervised learning methods to solve this problem. In its simplest form, supervised learning attempts to automatically learn a program from training data (Domingos, 2012). Supervised learning generally proceeds as follows. Annotated training data in the form of features (extracted from signals recorded by sensors as noted above) with temporally synchronized annotations of mental states (annotations are usually provided by humans) is collected as (ideally) a large number of people are in the midst of experiencing the mental states. Supervised learning methods are then applied to automatically *model* (learn) the relationship between the features and corresponding annotations (dashed lines in Figure 5.1). Aspects of the situational context are sometimes used as additional input to provide contextual information to the learning process. The resulting *model*, created during supervised learning, is then used to produce computer-generated estimates of mental states when presented with new data without corresponding annotations (e.g., collected at some later time and/or from a person not in the training data – solid lines in Figure 5.1). Accuracy and generalizability are the two immediate metrics of performance. Accuracy is measured as the extent to which automated mental state estimates align with some objective standard, typically self- or observer-reports of the mental states. Generalizability is concerned with the robustness of the detectors when applied to data beyond what was used to train the supervised classifiers.

It should be noted that the aforementioned discussion intentionally glosses over several of the complexities involved in the various stages of building an automated mental state detector. Data collection and annotation has subtle nuances that need to be mastered. Computing diagnostic features requires solving all sorts of open problems in respective fields (e.g., computer vision, acoustic signal processing, digital signal processing, time series modeling, natural language understanding). Then there is the issue of selecting a subset of diagnostic features, modeling relationships among features, and reducing dimensionality of the feature subspace. The training data usually need to be sampled and manipulated in numerous ways before supervised learning can proceed. Then comes the choice of supervised learning method from the multiple available possibilities (e.g., neural networks, Bayesian classifiers, decision trees, and support vector machines), followed by methods to parameterize the model. If multiple modalities (e.g., audio + visual) are used, then this brings about the additional challenge of deciding how to combine modalities. Finally, appropriate validation methods and metrics need to be selected, which is a nontrivial issue as well. Taken together, these challenges have supported rich and productive interdisciplinary research agendas and will continue to do so for years to come.

EXAMPLE SYSTEMS

This section turns to a discussion of a few exemplary case studies depicting the measurement of the variety of mental states of relevance to mental health care. The exposition is far from comprehensive, nor does it intend to be, since an in-depth discussion of the relevant work in this area would be more in line with an edited volume devoted solely to the subject of mental state detection. The idea is to highlight a few recent systems that have had some success in detecting mental states of particular relevance to mental health care.

Affective States

Affect is implicated in a number of mental disorders (e.g., depression, stress), so automated affect detection might be a promising way to obtain an indirect estimate of a person's underlying mental health. Affect is used as a general term to encompass both moods and emotions, which can be differentiated along a number of dimensions (Rosenberg, 1998). Moods are considered to be more transitory and have a background influence on consciousness, while emotions are brief, intense states that occupy the forefront of consciousness, have significant physiological and behavioral manifestations, and rapidly prepare the bodily systems for action. Most of the work on affect detection has focused on detecting emotions rather than moods, and an overwhelming majority of this work has emphasized the so-called "basic emotions," which typically include anger, surprise, happiness, disgust, sadness, and fear (Ekman, 1992). Nonbasic emotions, such as boredom, confusion, frustration, engagement, and curiosity, share a subset of the features commonly attributed to basic emotions (Ekman, 1992) and have received far less attention (D'Mello & Calvo, 2013). Some researchers eschew discrete affect representations (e.g., sad vs angry) by focusing on identifying levels of intensity on one or more core affect dimensions (Fontaine, Scherer, Roesch, & Ellsworth, 2007), with a specific emphasis on valence (unpleasant to pleasant) and arousal (sleepy to active).

Affect detection is one of the most widely studied mental state detection problems, evidenced by numerous recent reviews (Calvo & D'Mello, 2010; Calvo, D'Mello, Gratch, & Kappas, 2014; D'Mello & Kory, 2015; Zeng, Pantic, Roisman, & Huang, 2009). Affect detection systems greatly differ in terms of sensors/signals used, the affect representation (i.e., discrete vs dimensional), the specific affective states detected, whether the states naturally occur or are experimentally induced, and the contexts in which affect detection is situated. To give a sense of the work in this area, three affect detection efforts are reviewed, each emphasizing a different combination of sensor/signal, affect representation, affect state, and situational context.

Basic Emotions

The first study reviewed is a lab study that focused on detection of basic emotions elicited through an affect elicitation procedure. Janssen et al. (2013) compared automatic detection versus human perception of three basic emotions (happy, sad, angry), relaxed, and neutral induced via an autobiographical recall procedure (Baker & Guttfreund, 1993). According to this procedure, 17 participants were asked to write about two events in their life associated with experiences of happiness, anger, sadness, and neutral. Participants were then asked to recall a subset of those events in a way that made them relive the emotions experienced and to verbally describe each event (in Dutch). Audio, video, and physiological signals (electrodermal activity, skin temperature, respiration, and electrocardiography) were recorded while participants recalled and described the events. Each recording was associated with the label of the corresponding emotion that the participant was asked to recall.

A variety of standard features, such as specific facial landmarks, head position, fundamental frequency of speech, and overall level and variance in each physiological signal, were automatically computed. A support vector machine classifier (supervised learning method) produced the best result when it only focused on facial and physiological features. It obtained an accuracy of 82% in correctly identifying the emotion label of each recording. In addition, the authors directly compared computer detection to human detection of emotion. This was done by asking a set of human judges (both US and Dutch) to identify the participants' emotions based on various stimuli combinations (audio-only, video-only, audio-video). The Dutch judges were the most accurate (63%) when only provided with the audio (which was also in Dutch) while the US judges were the most accurate (31%) when they had access to both audio and video. However, accuracy of the humans (63% and 31%) was considerably lower than accuracy of the automated detector (82%), a finding with profound implications.

Nonbasic Emotions

The second study (Bosch et al., 2015) adopted a markedly different approach from Janssen et al. (2013), which focused on multimodal detection of (mostly) basic emotions experimentally elicited in controlled laboratory environments. Bosch et al. (2015) studied unimodal detection of nonbasic emotions that naturally occurred in a noisy real-world setting of a computer-enabled classroom. In this study, 137 middle and high school students played a conceptual physics educational game in small groups for 2.5 h across 3 days as part of their regular physics/physical science classes. Trained observers performed live affect annotations by observing students one at a time using a round-robin technique (observing one student until visible affect was detected or 20 s had elapsed and then moving on to the next student in a preplanned order – see Ocumpaugh, Baker, and Rodrigo (2012)). The emotions of interest were boredom, confusion, delight, engagement, and frustration. Videos of students' faces and upper bodies were recorded during game-play and synchronized with the affect annotations. The videos were processed using the FACET computer-vision program (Emotient, 2014), which provides estimates of the likelihood of 19 facial AUs (Ekman & Friesen, 1978) (e.g., raised brow, tightened lips), head pose (orientation), and position. Body movement was also estimated from the videos using motion-filtering algorithms (Kory, D'Mello, & Olney, in press). A machine learning approach was adopted to automatically discriminate each affective state from all the others and was validated in a manner that generalizes to new students. Person-independent automatic detection accuracies ranged from 62% (frustrated vs. other states) to 83% (delighted vs. other states), which is notable given the noisy nature of the environment with students incessantly fidgeting, talking with one another, asking questions, leaving to go to the bathroom, and even occasionally using their cellphones (against classroom policy).

Affect Dimensions

The third study reviewed is not exactly a study, but a collection of different efforts aimed at solving a particular affect detection problem. The idea is that it is difficult to ascertain progress in any given research area (affect detection in this case) when individual researchers apply their own methods to their own data sets and use their own selected metrics to evaluate performance. Direct comparisons of results from different research groups are confounded as any observable difference can be attributed to the method, the data, or the performance metric. Challenge competitions, a common theme in computer science and AI research, offer one answer to this problem. Here, researchers are asked to apply their methods to a fixed dataset and the results are evaluated with a fixed metric(s), thereby affording direct comparisons across methods developed by different research groups.

The Audio-Video Emotion Recognition Challenge (AVEC) is an annual affect detection challenge that was first organized as part of the 2011 Affective Computing and Intelligent Interaction (ACII) conference series (D'Mello, Graesser, Schuller, & Martin, 2011). The focus here is on the 2012 AVEC challenge (Schuller, Valster, Eyben, Cowie, & Pantic, 2012), which considered automatic detection of affect dimensions during human–computer interactions. The AVEC 2012 challenge used data from the SEMAINE corpus (McKeown et al., 2012), which was designed to collect naturalistic data of humans interacting with artificial agents. The artificial agents take on different emotionally stereotyped roles (e.g., Spike is angry and confrontational while Prudence is even-tempered and sensible), thereby biasing the affective tone of the conversation. Videos of participants' faces and audio of their speech recorded during these somewhat affectively charged interactions were provided to researchers. Two to six human raters annotated each video along four affect dimensions: valence (negative to positive), arousal (sleepy to active), power (low control to high control), and expectation (unexpected to expected). The affect annotations were continuously scaled and ranged from −1 to 1, so the task was to predict the intensity of each affect dimension via automated audio-visual affect detection methods. This emphasis on dimensional representations of affect (e.g., valence, arousal, power, expectation) is a more important discriminating factor than the categorical or discrete representations (e.g., anger, fear) adopted in the previous two studies reviewed.

Researchers were given two subsets of the annotated data to develop their models (training and development subset), which were then applied on a separate subset for which the annotations were not available to the researchers (test subset). Each research group submitted affect predictions for each dimension independently by applying their methods on the test subset. Results were available from 10 research groups. The winning team achieved a correlation of 0.45 when averaged across the four dimensions (Nicolle, Rapp, Bailly, Prevost, & Chetouani, 2012), a notable performance given the complexity of the task.

Attentional Lapses (Mind Wandering)

Mindfulness, or the ability to devote one's attentional resources to the present task and surroundings, is considered to be an important component of mental health (Brown & Ryan, 2003). However, mind wandering, defined as involuntary lapses in attention from the task at hand to internal task-unrelated thoughts (Smallwood & Schooler, 2015), is all too frequent. For example, in a large-scale study, mind wandering was tracked in 5,000 people from 83 countries working in 86 occupations with an iPhone app that prompted people to report off-task thoughts at random intervals (Killingsworth & Gilbert, 2010). The main finding was that people reported mind wandering for 46.9% of the prompts, and a time-lagged analysis provided some evidence that mind wandering could potentially cause feelings of unhappiness. Furthermore, research has indicated that individuals with dysphoria (i.e., symptoms of depression) and ADHD have elevated levels of mind wandering across a variety of tasks (Shaw & Giambra, 1993; Smallwood et al., 2004; Smallwood, O'Connor, Sudbery, & Obonsawin, 2007). Thus, automated detectors of mind wandering as people go about their daily routines have the potential to provide valuable information into the attentional processes implicated in mental health.

Automated detectors of mind wandering are just beginning to emerge. As is typical for early stages of research, initial efforts are concentrated on one or more restricted task domains. In the case of mind wandering, the focus has been on reading comprehension tasks (Bixler & D'Mello, 2014; Blanchard, Bixler, & D'Mello, 2014; Drummond & Litman, 2010; Franklin, Smallwood, & Schooler, 2011). As an illustrative example, Bixler and D'Mello (2014) investigated whether eye gaze could be used to develop an automated mind wandering detector during reading. Their focus on eye gaze was motivated by decades of research highlighting the tight coupling between visual attention and eye movements (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Rayner, 1998). In their study, training data were collected from 178 undergraduate students from two US universities in a lab study. Tobii T60 and TX 300 eye trackers (one at each university) were used to record gaze patterns of the students over the course of a 30-min text-comprehension activity. Mind wandering was measured via auditory thought probes, which is a standard technique used for online tracking of mind wandering (Smallwood et al., 2004). Eye gaze features were computed from gaze fixations in windows of variable length (3, 5 s, etc.) that immediately preceded a mind wandering probe. Example features include fixation durations, number of words skipped, fixations on different types of words, and saccade lengths. Supervised learning methods, applied to detect mind wandering, yielded an average recognition rate of 66% and in a manner that generalized to new people. Though preliminary, these results are promising since they denote the possibility of automatically measuring a highly internal mental state like mind wandering. The increased availability of cost-effective consumer-grade eye trackers (some retail for as low as $99) also suggests that this line of work can soon be taken out of the lab and into the wild.

Pain

Pain has been associated with numerous mental disorders. A large-scale study of 85,088 people from 17 countries indicated that chronic back/neck pain in a 12-month period was a positive predictor of mood disorders, anxiety disorders, and alcohol abuse/dependence after covarying age and sex (Demyttenaere et al., 2007). Measurement of pain, which primarily relies on self-report questionnaires (Hjermstad et al., 2011), is subject to well-known limitations of subjectivity, interpretability, and feasibility of administration in some populations (Stinson, Kavanagh, Yamada, Gill, & Stevens, 2006).

Automatic pain detectors can alleviate several of these challenges by offering reliable monitoring of pain. The numbers of automated pain detection systems are few and far between (Hammal & Cohn, 2012; Lai, Levinger, Begg, Gilleard, & Palaniswami, 2009; Littlewort, Bartlett, & Lee, 2007), presumably due to the complexities in obtaining suitable datasets for detector building. However, the recent release of the UNBC-McMaster shoulder pain expression archive database is expected to catalyze research in this area. The database includes 129 participants who self-identified as experiencing shoulder pain. The data consist of videos of participants performing eight range of motion tests (e.g., abduction, internal and external arm rotation) along with self-reports of pain intensity following each test. A subset of the data (200 video sequences from 25 participants) has been made available to the research community for the purpose of building automatic pain detection systems (Lucey, Cohn, Prkachin, Solomon, & Matthews, 2011).

Hammal and Cohn (2012) provide an illustrative example of one such automated pain detection system utilizing the UNBC-McMaster database. Their approach consisted of extracting appearance-based features from each frame in the video and filtering them via a set of log-normal filters. A support vector machine classifier was used to build detectors of four levels of pain (no pain, trace pain, weak pain, strong pain). They achieved an average classification accuracy of 0.56 (F1 metric) when the validation method ensured generalizability to new participants, a promising result given the complexity of the problem and the early stages of research in this area.

Depression

Depression is perhaps one of the most common and serious mental health conditions. Automatic depression detection systems have considerable potential to combat its negative effects by providing early warning indicators of depression as well as serving as an objective measure of the effectiveness of depression treatments. Research in depression detection has recently accelerated, presumably due to the introduction of the Depression Recognition Sub-challenge (DSC) as part of the 2013 and 2014 AVEC series (Valstar et al., 2013, 2014). The challenge requires researchers to develop and evaluate their own depression detectors on the same dataset, thereby affording meaningful comparisons of each method since data and evaluation metrics are held constant.

The dataset used in the DSC challenge consisted of 240 h of videos (with audio) of 84 participants who completed simple tasks guided by a computer interface across multiple sessions. The number of tasks, sessions, and session length varied across participants. Participants' levels of depression were obtained via the Beck Depression Inventory-II (Beck, Steer, & Brown, 1996). The data were also annotated for basic affect dimensions, but these are not discussed here. A subset of these data was used for the 2013 and 2014 challenges. In the most recent 2014 challenge, this included 300 videos of participants reading aloud in German excerpts from a German fable (northwind task) and responding to simple questions in German (e.g., "what is your favorite dish" – freeform task). Videos of participants' faces and audio of their speech were recorded during these tasks. A subset of these data along with depression levels of each participant were made available to researchers (training and development partitions). A different subset, called the test partition, was used to evaluate results and the depression levels for these participants were withheld.

Researchers adopted a wide variety of approaches in response to this challenge. The results were quantified via root mean square error (RMSE) between predicted and actual depression levels for participants in the test partition and ranged from 8 to 12. The winning system emphasized modeling the timing and coordination between speech production and facial expression and achieved an RMSE of 8.12 (Williamson, Quatieri, Helfer, Ciccarelli, & Mehta, 2014). This result represented a small improvement over the best result of the 2013 challenge (RMSE of 8.50), which was won by the same research team (Williamson et al., 2013) on a related but different dataset.

Stress

Both episodic and chronic stress have long been associated with numerous physical and mental health outcomes (see review by Lupien, McEwen, Gunnar, & Heim, 2009). Automatic detection of stress has important applications for stress diagnosis, treatment, and monitoring. Consequently, considerable research has been devoted toward the design of automatic stress detection systems. Computational researchers use the term stress somewhat broadly to include cognitive stress (e.g., induced by taxing working memory), emotional stress (e.g., induced via negative activating emotions), and social stress (e.g., induced by social stressors). The focus of this chapter is on automatic detection of clinical stress as measured via the widely used Perceived Stress Scale (PSS) (Cohen, Kamarck, & Mermelstein, 1983).

The study by Sano and Picard (2013) is unique from the other studies reviewed in this chapter because it adopts a multimodal approach for ambulatory stress monitoring of stress across a period of 5 days. Eighteen participants were asked to complete the PSS along with other pertinent self-report questionnaires (sleep habits, alcoholic consumption, moods). Participants were then affixed with an Affectiva Q-sensor, a wearable wristband with an electrodermal activity sensor and an accelerometer. An Android application installed on their phones was used to record various aspects of their daily routine including phone usage (e.g., number of calls made, texts received, screen on times) as well as location using GPS (e.g., distance traveled). Participants also completed a short morning survey and an evening survey on their phones (e.g., sleep time, alertness on waking up, cups of coffee consumed in the day). The physiological, activity, and survey data were collected for a period of 5 days as participants went about their everyday activities. Participants were then assigned to a low-versus high-stress group based on their scores on the PSS. Supervised learning methods were then used to discriminate among the low- and high-stress groups using various combinations of features. As a baseline, the authors reported a classification accuracy of 87.5%, obtained via the use of the self-report surveys. However, accuracies of approximately 75% were obtained using solely objective information gleaned from the activity logs based on cell phone usage and location data. Surprisingly, the physiological features did not make a substantial contribution to stress detection on these data. Nevertheless, the fact that stress could be predicted with reasonable accuracy from activity logs and location data is a significant finding because it highlights the potential for ambulatory stress monitoring using an everyday device (a smart phone) and without requiring any additional sensors.

CONCLUDING REMARKS

Measurement is a precursor to meaningful change. The science and practice of mental health care has much to gain from fully automated systems that provide fine-grained assessments of a person's mental state over extended periods of time and in a variety of contexts. These mental state detection systems can be integrated within the overall mental healthcare system at multiple levels, for example clinical decision-making, ambulatory monitoring, and technology-supported therapies. This chapter discussed some of the theoretical and technical issues underlying such systems and grounded the key issues in the context of a few case studies focused on automatically detecting mental states of relevance to mental health care (i.e., affect, attention (or lack thereof), pain, depression, and stress). The highly selective nature of this review regrettably precluded a discussion of many other excellent systems developed by dedicated groups of international researchers in many different fields who are continually making theoretical, technical, and practical innovations to crack the challenge of automatic mental state detection.

Automatic mental state detection is a tough nut to crack. The current systems are not yet ready for practical use although there has been much progress over the years. Many of the earlier years in the field were focused on demonstrating research prototypes as proof-of-concepts of the possibility of automatic mental state detection. This was necessary to convince the initial skeptics and naysayers who ridiculed the early pioneers of the field (as discussed in Picard, 2010). These early (generation 1) systems focused on a small subset of mental states that were acted (or induced) by a small number of people in the confines of the laboratory. Generation 1 was also marked by the use of expensive and obtrusive sensing devices that were inherently nonscalable and with the use of less technically sophisticated computational techniques and less stringent validation methods. We are now in generation 2, where the emphasis is on detecting naturalistic experiences of a larger variety of mental states in more real-world contexts using more scalable, wearable, and unobtrusive sensing and with more sophisticated techniques and more stringent validation methods. Much progress is anticipated for these generation 2 systems, but they may still fall short in some respects. In particular, there needs to be an eye for improving detection accuracies, for demonstrating applicability across a range of real-world contexts, for realizing generalizability across different populations, and for satisfactorily addressing thorny ethical issues. It is not a matter of “will” but “when” these challenges will be addressed, upon which automatic mental state detection systems will make a meaningful and measurable impact on peoples’ lives by improving their mental health.

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) (DRL 1108845) and the Bill & Melinda Gates Foundation. Any opinions, findings and conclusions, or recommendations expressed in this chapter are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

Baker, R. C., & Guttfreund, D. O. (1993). The effects of written autobiographical recollection induction procedures on mood. *Journal of Clinical Psychology*, *49*(4), 563–568.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–645.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck depression inventory-II*. San Antonio, TX: Psychological Corporation.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, *78*(Suppl. 1), B231–B244.

Bixler, R., & D’Mello, S. (2014). Toward fully automated person-independent detection of mind wandering. In V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, & G.-J. Houben (Eds.), *Proceedings of the 22nd international conference on user modeling, adaptation, and personalization* (pp. 37–48). Switzerland: Springer International Publishing.

Blanchard, N., Bixler, R., & D’Mello, S. K. (2014). Automated physiological-based detection of mind wandering during learning. In S. Trausan-Matu, K. Boyer, M. Crosby, & K. Panourgia (Eds.), *Proceedings of the 12th international conference on Intelligent Tutoring Systems (ITS 2014)* (pp. 55–60). Switzerland: Springer International Publishing.

Bosch, N., D’Mello, S. K., Baker, R., Ocumpaugh, J., Shute, V., Ventura, M., et al. (2015). Automatic detection of learning-centered affective states in the wild. In *Proceedings of the 2015 international conference on Intelligent User Interfaces (IUI 2015)* (pp. 379–388). New York, NY: ACM.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822.

Calvo, R. A., & D’Mello, S. K. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37. Available from: http://dx.doi.org/10.1109/T-AFFC.2010.1.

Calvo, R. A., D’Mello, S. K., Gratch, J., & Kappas, A. (Eds.), (2014). *The Oxford handbook of affective computing*. New York, NY: Oxford University Press.

Coan, J. A. (2010). Emergent ghosts of the emotion machine. *Emotion Review*, *2*(3), 274–285.