X

Multimodal-Multisensor Affect Detection

Sidney K. D'Mello[1], Nigel Bosch[2], and Huili Chen[3]

[1]University of Colorado Boulder

[2]University of Illinois Urbana-Champaign

[3]Massachusetts Institute of Technology

Imagine you are interested in analyzing the emotional responses of a person in some interaction context (i.e., with computer software, a robot, in a classroom, on the subway). You could simply ask the person to self-report his or her felt emotion using a questionnaire, a valence-arousal grid [Russell et al. 1989], a self-assessment manikin [Bradley and Lang 1994], or some such measurement instrument. Or you could ask trained humans to observe the person and provide emotion judgments [Ocumpaugh et al. 2015]. You could also record audio/video of the interaction and have trained coders annotate the videos for visible emotion at some later time. You can even use computer vision techniques to obtain automatic estimates of facial expressions in the videos [Girard et al. 2015]. Or you may be interested in the person's physiological responses and can use a variety of sensors to collect these data.

These examples capture some (but not all) of the contemporary approaches to measure emotional responses [Coan and Allen 2007]. The approaches can be categorized as subjective vs. objective, each with different affordances. The subjective approaches (self and observers) are *best* suited for emotion-level representations (e.g., discrete emotions like anger and fear or dimensional representations like valence or dominance) at coarse-grained temporal resolutions (tens of seconds to minutes). The objective approaches (sensors and software) are *ideal* for measurement of behavioral/physiological responses (e.g., facial expressions, electrodermal activity) at fine-grained temporal resolutions (milliseconds to seconds). The two approaches have complementary strengths and weaknesses. The subjective approaches capitalize on humans' knowledge and reasoning capabilities, resulting in more nuanced and contextualized emotion assessments. However, they are limited by fatigue, biases (e.g., social desirability bias), errors (e.g., memory reconstruction for self-reports), and are difficult to scale. The objective approaches are not affected by fatigue or biases and are more scalable, but have limited inference and reasoning capabilities, thereby mainly providing readouts of behavioral/physiological responses.

Are there ways to reconcile the two approaches? One strategy is to combine both, for example, collecting subjective self-reports of frustration in tandem with computerized estimates of facial action units (AUs) [Girard, Cohn, Jeni, Sayette and De la Torre 2015]. The two are taken as complementary perspectives of the person's emotional response and associations are analyzed offline, i.e., by correlating self-reports of frustration with AUs. But what if there was a way to combine both perspectives on the fly so that the measurement jointly reflects both subjective emotion perception by humans and objective behavioral/physiological signals recorded by sensors? And what if the measurement could occur in a fully automated fashion, thereby providing

measurement at fine-grained temporal resolutions and at scale? And further, what if the measurement engine was sufficiently sophisticated to model multiple expressive channels and the nonlinear temporal dependencies among them? This is the affective computing (AC) approach to emotion measurement and is the focus of this chapter.

Affective computing [Calvo et al. 2015; Picard 1997], broadly defined as computing involving or arising from human emotion, is an interdisciplinary field that integrates the affective and computational sciences. Affect detection (or affect recognition) is one of the key subfields of affective computing (see reviews- [Calvo and D'Mello 2010; D'Mello and Kory 2015; Zeng et al. 2009]). The goal of affect detection is to automatically provide estimates of latent higher-level affective representations (e.g., fear) from machine-readable lower-level response signals (e.g., video, audio, physiology). Multimodal-multisensor affect detection (MMAD) utilizes multiple modalities (e.g., video, cardiac activity) and/or multiple sensors (e.g., video, electromyography) as an alternative to unimodal affect detection (UMAD).

In this chapter, we provide a conceptual and technical overview of the field of MMAD, ground the abstract ideas via walk-throughs of three MMAD systems, and provide a summative review of the state-of-the-art in the field. We begin with a background discussion from the affective sciences, starting with a very basic question: "what is affect?"

Table 1. Key terms with operational definitions

Term	Operational definition
Construct	A conceptual variable that cannot be directly observed (e.g., intelligence, personality)
Affect	Broad term encompassing constructs such as emotions, moods, and feelings. Is not the same as personality, motivation, and other related terms.
Affective experience-expression link	The relationship between experiencing an affective state (e.g., feeling confused) and expressing it (e.g., displaying a furrowed brow).
Affect annotation	The process of assigning affective labels (e.g., bored, confused, aroused) or values (e.g., arousal = 5) to data (e.g., video, audio, text)
Affective ground truth	Objective reality involving the "true" affective state. Is a misleading term for psychological constructs like affect
Affective computing	Computing techniques and applications involving emotion or affect
Multimodal fusion	The process of combining information from multiple modalities
User-independent model	A model that generalizes to a different set of users beyond those used to develop the model

X.1 Background from affective sciences

Affect

What is affect? The simple answer is that affect has something to do with feeling. Perhaps a more satisfactory answer is that affect is a broad label for a range of psychological phenomena involving feelings. This includes primitive feelings like hunger pangs to more complex social emotions like jealousy and pride. A more technical answer is that affect is a multicomponential construct (i.e., conceptual entity), that operates across neurobiological, physiological, behavioral, cognitive, metacognitive, and phenomenological levels [Barrett 2014; Lewis 2005; Mesquita and Boiger 2014; Scherer 2009]. It is with good reason that none of these answers seem particularly satisfactory. The term affect (or emotion) has resisted attempts at crisp definition despite a century of concentrated effort [Izard 2010; Izard 2007]. Understanding what emotions are and how they arise has been a contentious issue in the affective sciences and is sometimes referred to as the "hundred year emotion war" [Lench et al. 2013; Lindquist et al. 2013]. For example, there has been an ongoing debate as to whether affect is best represented via discrete categories (e.g., angry, fearful) [Lerner and Keltner 2000; Loewenstein and Lerner 2003] or by fundamental dimensions (e.g., valence, arousal, power) [Cowie et al. 2012; Russell 2003] (and on how many dimensions are needed [Fontaine et al. 2007]). Other open issues pertain to whether emotions are innate or are learned, whether they arise via appraisals/reappraisals or are they products of socio-constructivism, and whether emotions are universally expressed or if context and culture shape emotion expression [Barrett 2006; Barrett et al. 2007; Ekman 1992; Ekman 1994; Gross and Barrett 2011; Izard 1994; Izard 2010].

Does the fact that we cannot precisely define affect imply that we cannot detect it? In our view, one does not need to precisely define a phenomenon in order to study it. However, researchers need to be mindful of the implicit assumptions in their operationalizations of affect as these are transferred to the affect detectors. For example, if one operationalizes anger as short-term emotional changes recorded while people viewing anger-eliciting films in isolation and builds an automated anger detector from these recordings, then the detector's estimates of anger are inherently coupled to this precise operationalization and not much else (e.g., felt anger, anger in a road-rage scenario, anger in a social context). Thus, it is important to be mindful that measurement is informed by assumptions of reality (operationalizations), which, in turn, are informed by insights gleaned by measurement.

The affective experience-expression link

Affect detection assumes a link between experienced (or felt) and expressed affect. Thus, it should be theoretically possible to "decode" latent affect (e.g., confusion) from visible behaviors (e.g., a furrowed brow). This suggests that there exist "mappings" between a set of behaviors (e.g., facial features, gestures, speech patterns) and a set of affective states. This does not mean that one simply needs to learn the mappings to perfectly solve the affect detection problem because the mappings are imprecise. For example, although facial expressions are considered to be strongly associated with affective states, meta-analyses on correlations between facial expressions and affect have yielded small to medium effects under naturalistic conditions [Camras and Shutter 2010; Fridlund et al. 1987; Ruch 1995; Russell et al. 2003]. In the interest of maximizing adaptability to new situations and environments, the mappings have evolved to be loose and variable, not fixed and rigid [Coan 2010; Roseman 2011; Tracy 2014]. Thus, rather than being predefined, the affect-

expression links emerge from dynamic interactions between internal processes and the environmental context. Some of these influences include the internal state of the individual, contextual and social factors [Parkinson et al. 2004], and individual and group (or cultural) differences [Elfenbein and Ambady 2002a; Elfenbein and Ambady 2002b].

At first blush, the lack of a precise experience-expression link seems to threaten the entire affect detection endeavor. But this is not the case. In our view, it is sufficient to assume that there is *some* link between experience and expression. The link need not be particularly strong. The link need not even be consistent across individuals, situations, and cultures. The only assumption is that there is a "beyond-chance probabilistic" [Roseman 2011 p., 440] link between affect expression and experience. Most affect detection systems rely on supervised learning methods to learn this link. Supervised learning needs supervision in the form of "ground truth" (annotations) which bring us to the question of "what is affective ground truth?"

Affective ground truth

Consider speech recognition, where the task is to translate an acoustic representation into a linguistic representation of speech. There is usually little dispute about the desired output (i.e., the words being spoken). But this is rarely the case with affect detection as affect is a psychological construct (see above). One exception is when the affective states are portrayed by actors or are experimentally induced [Kory and D'Mello 2015]. Here, the acted/induced affect can be taken as ground truth, but the resultant expressions more closely resemble the acting/eliciting micro-context and might not generalize more broadly (also see [André 2017] – this volume).

There is no objective ground truth in the case of naturally occurring affective states. Instead, the truth lies in the eyes of the beholder. The beholder, in the case of humans, is the person experiencing the emotion (the self) or an external observer. Each has access to different sources of information and is subject to different biases, thereby arriving at different approximations of "ground truth." As noted above, affective states are multicomponential in that they encompass conscious feelings ("I feel afraid"), overt actions ("I freeze"), physiological/behavioral responses ("My muscles clench"), and meta-cognitive reflections ("I am a coward"). Access to these components varies by source (self vs. observer). The self has access to some conscious feelings, some overt actions, memories of the experience, and meta-cognitive reflections, but usually not to some of the unconscious affective components. They are also more likely to distort or misrepresent their affective states due to biases, such as reference bias [Heine et al. 2002] or social desirability bias [Krosnick 1999]. In contrast, observers only have access to overt actions and behaviors that can be visibly perceived (e.g., facial features, postures, gestures) and must rely more heavily on inference [Mehu and Scherer 2012]. Observers are less likely to succumb to the same biases that befall self-reports, but they introduce biases of their own, such as the halo effect [Podsakoff et al. 2003]. There are strengths and pitfalls of reliance on either the self or external observers to establish affective "ground truth." [D'Mello 2016]. Therefore, perhaps the most defensible position is to consider a combination of perspectives, thereby capitalizing on their merits while minimizing their flaws.

Multimodal coordination of affective responses

Consider the following quote from William James in his classic 1884 treatise, "What is an emotion?"

> "Can one fancy the state of rage and picture no ebullition of it in the chest, no flushing of the face, no dilatation of the nostrils, no clenching of the teeth, no

> impulse to vigorous action, but in their stead limp muscles, calm breathing, and a placid face?" [James 1884 p. 452]

Quotes such as the one above by James [1884] and similar ones by Darwin [1872], Tomkins [1962], Ekman [1992], Damasio [2003] and others, depict affective responses as being inherently multimodal. According to the classical model of emotion (called basic emotion theory), there is a specialized circuit for each (basic) emotion in the brain. Upon activation, this circuit triggers a host of *coordinated responses* encompassing peripheral physiology, facial expression, speech, modulations of posture, affective speech, instrumental action, cognitions, and subjective experience [Ekman 1992; Izard 2007]. According to this view, MMAD should be substantially more accurate than UMAD because MMAD approaches model this coordinated emotional response.

In contrast to this highly integrated, tightly coupled, central executive view of emotion, researchers have recently argued in favor of a disparate, loosely coupled, distributed perspective [Coan 2010; Lewis 2005]. Here, there is no central affect neural circuit [Lindquist et al. 2016; Lindquist et al. 2011] that coordinates the various components of an emotional episode. Instead, these components are *loosely coupled* and the situational context and appraisals determine which bodily systems are activated and the dynamics of activation over time. These theories would accommodate the prediction that that a combination of modalities might conceivably yield small improvements in classification accuracies, suggesting that the merits of MMAD over UMAD approaches might not necessarily lie in improved classification accuracy, but in other factors (e.g., increased reliability due to redundancy).

We consider the extent the data supports each of these views later on in the chapter. The reader is also directed to Vinciarelli and Esposito [2017] (this volume) for a discussion on the conditions when multimodal communication should expect benefits over unimodal signaling. There is also a parallel line of work focused on human perception of affect from unimodal and multimodal cues expressed by both humans [D'Mello et al. 2013] and virtual agents [Martin et al. 2017] (this volume), that could establish baselines for what machines might be capable of achieving.

X.1 Modality fusion for multimodal-multisensor affect detection

Figure 1 highlights our theoretical position on affective states (see previous section), which informs the steps involved in building an affect detector. Affective states are assumed to emerge from person-environment interactions and are reflected in changes at multiple levels (i.e., neurobiological changes, physiological responses, bodily expressions, action tendencies, and cognitive, metacognitive, and phenomenological states) in a manner that is modulated by individual differences (e.g., affective traits, culture). Researchers typically adopt a machine learning approach for affect detection, which requires the collection of training and validation data. Accordingly, in Step 1a, raw signals (video, physiology, event log files, etc.) are recorded as participants engage in some interaction of interest (including experimental elicitation). Features are then computed from the raw signals (Step 1b). Affect annotations (Steps 2a and 2b) are obtained from the participants themselves or from external observers, either online (e.g., live observations) or offline (e.g., video coding). If affect is experimentally induced, then the elicited condition serves as the annotation. Next, machine learning methods (typically supervised learning) are used to computationally model the relationship between the features and the affect annotations (Step 3). The models can also include contextual information, including both external context (e.g., situational aspects, task

constraints, social environment) and internal context (e.g., previous affect predictions). The resulting machine-learned model yields computer-generated annotations, which are compared to the human-provided annotations in a validation step (Step 4). Once validated, the computational model can now produce computer-generated affect annotations from a new set of raw signals without corresponding human-provided annotations..

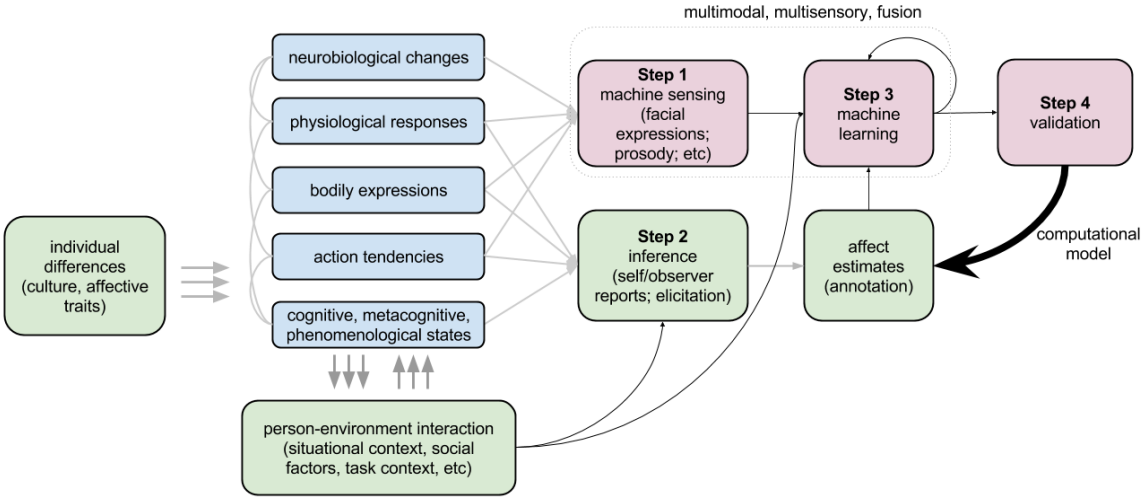

Figure 1. Theoretical foundation and steps involved in affect detection

The basic affect detection approach needs an update when multiple modalities and/or sensors are involved. They key issue pertains to how to synchronize and combine (fuse) the different information channels (modalities). In the remainder of this section, we explore a variety of methods for this task. Alternate fusion methods, specifically for online affect detection, are discussed in André [2017] (this volume).

Basic methods (data, feature, decision, and hybrid fusion)

The most basic method for fusing modalities is data-level or stream-level fusion. Here, raw signals are first fused before computing features. For example, one might record electrodermal activity (EDA) from multiple sensors to compensate for left-right EDA asymmetry [Picard et al. 2015] and then fuse the two signals (e.g., via convolution) prior to computing features.

The next basic method is feature-level fusion (or early fusion), where features from different modalities are concatenated prior to machine learning (see Figure 2). The primary advantage of feature-level fusion is its simplicity and it can be effective when features from individual modalities are independent and the temporal dependencies among modalities are minimal.

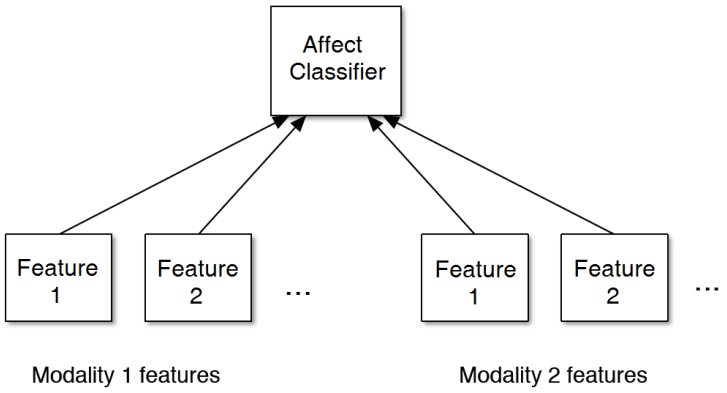

Figure 2. Illustration of feature-level fusion

An alternative is decision-level (or late fusion) fusion (Figure 3), where models are trained for each modality. The final decision is made by fusing the outputs of models corresponding to each modality via majority voting, weighting votes according to accuracy of each model, or training a new classifier using the outputs of each model as features (stacking).

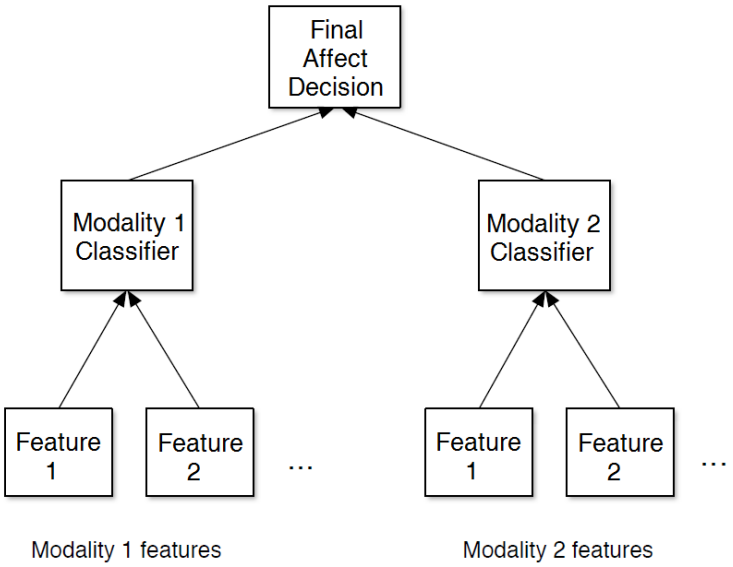

Figure 3. Decision-level fusion with two modalities

It is also possible to combine feature- and decision- level fusion as illustrated in Figure 4. The resultant method, called hybrid fusion, is expected to capitalize on the merits of each approach.

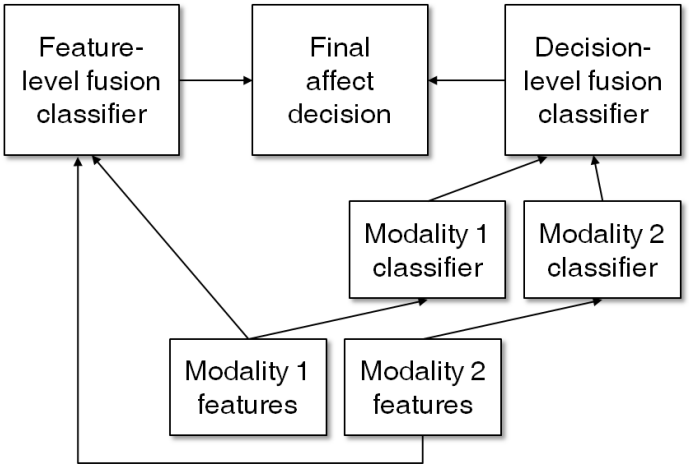

Figure 4. Hybrid fusion with two modalities

Model-based fusion with Dynamic Bayesian Networks (DBNs) and Hidden Markov Models (HMMs)

The aforementioned basic fusion methods are limited in that they do not account for temporal relationships among modalities. There are more sophisticated fusion methods, but these also ignore temporal dependencies. For example, in a support vector machine classifier, a kernel function is used to map input data to a higher-dimensional space. Multimodal fusion can be achieved by tuning a different kernel for each modality (feature space) and mapping them all into the same higher-dimensional feature space [Liu et al. 2014]. A limitation, however, is that these methods do not afford modeling of temporal dependencies, which is critical for MMAD. Model-based fusion methods model temporal dependencies as well as other relationships as illustrated with two widely used graphical models: Dynamic Bayesian Networks and Hidden Markov Models,

Dynamic Bayesian Networks (DBNs) are a common graphical model used for modality fusion in affect detection. Links between variables in DBNs represent conditional dependencies between features as well as relationships across time. Figure 5 shows a DBN that fuses two modalities along with contextual (top-down) features with *Affect* being the output variable. Top-down features (e.g., age; context factors) influence affect, but do not change from one timestep to the next. Bottom-up features, such as facial expressions and bodily movements, are linked across time. Affect also evolves across time, [D'Mello and Graesser 2011] so the *Affect* variable is linked across timesteps. Bayesian inference is used to compute the probability of the output *Affect* variable given the top-down (predictive) and bottom-up (diagnostic) features [Conati and Maclaren 2009].

DBNs have successfully been used in several MMAD systems. Li and Ji [2005] fused a variety of modalities including facial expressions, eye gaze, and top-down features (physical condition and time in circadian rhythm) to detect fatigue, nervousness, and confusion. Chen et al. [2009] detected anger, happiness, meanness, sadness, and neutral using a DBN to fuse audio and visual features. Jiang et al. [2011] expanded that work to detect a larger set of affective states including anger, disgust, fear, happiness, sadness, and surprise, using a similar DBN. In general, DBNs are quite flexible, allowing any structure of relationships between variables and across time. However, more complex DBN structures require considerably more training data to estimate the various parameters,

so in practice relatively simple structures like Figure 5 are used.

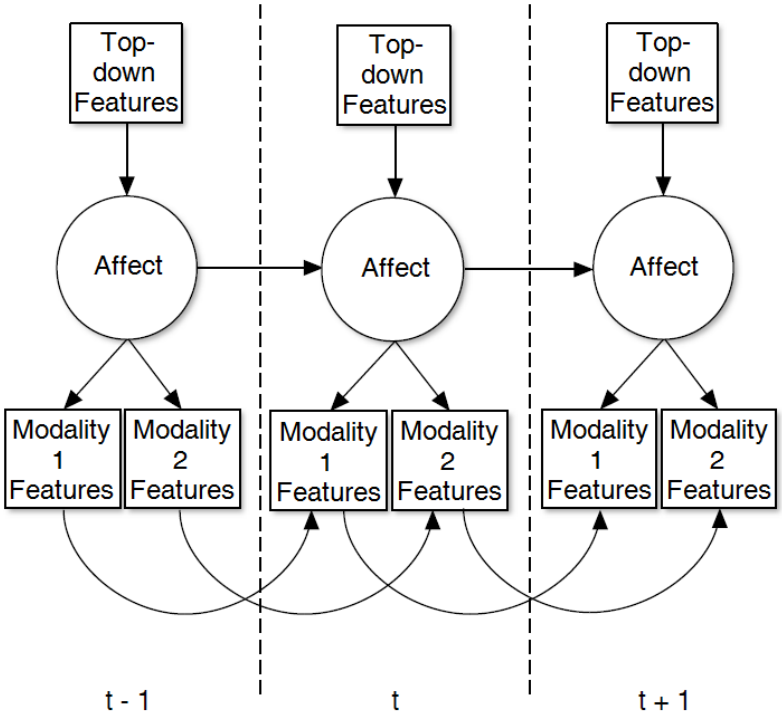

Figure 5. Dynamic Bayesian network model fusing two modalities and top-down features

One such structure is a Hidden Markov model (HMM), which models affect as a hidden variable that influences observable variables (e.g., anger influencing skin conductance and heart rate). Coupled hidden Markov models (CHMMs) combine two or more HMMs (one per modality), such that the hidden states (representing affect) of the individual HMMs interact across time (see Figure 6). These cross-modal links in a CHMM are chosen to model temporal relationships between modalities that might operate at different time scales (e.g., heart rate vs. facial expressions). As an example, Lu and Jia [2012] used a CHMM to combine audio and video HMMs to detect affect represented in an evaluation-activation (valence-arousal) space.

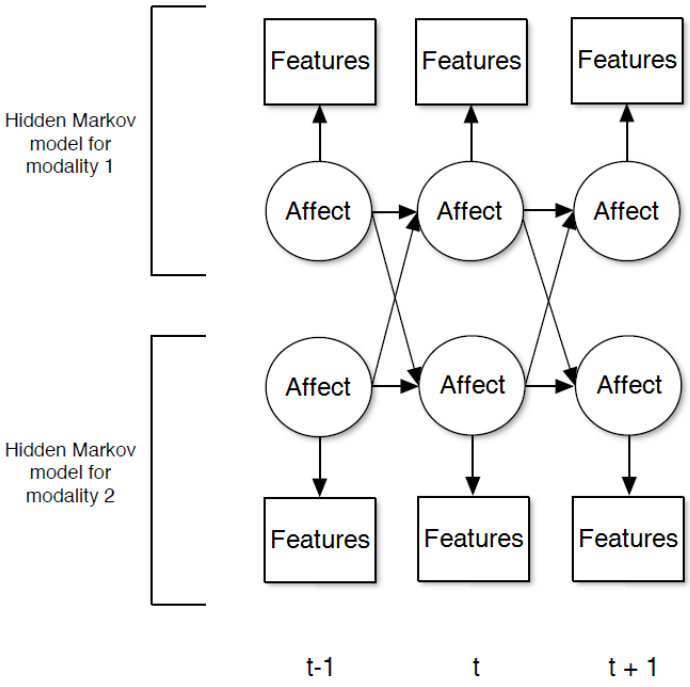

Figure 6. Coupled hidden Markov model for two modalities

CHMMs capture the temporal relationships between modalities, but consider each modality as a whole. Semi-coupled Hidden Markov models (SCHMMs) extend the structure of CHMMs by coupling modalities at the feature level. Component models are created for each pair of features, resulting in a large number of small models which are subsequently combined by late fusion. The main advantage of the SCHMM approach is that it allows the temporal relationships to vary per feature pair. Lin et al. [2012] demonstrated that SCHMMs were effective for recognizing affect on two audio-visual datasets, one with evaluation-activation dimensions and one with anger, happiness, sadness, and neutral. They found that SCHMMs outperformed standard CHMMs on both datasets.

Modality fusion with neural networks and deep learning

Neural networks have emerged as another popular approach for modality fusion. One particularly prominent type of network is the long short-term memory (LSTM) neural network [Hochreiter and Schmidhuber 1997]. In LSTMs, the artificial neurons in the hidden layers are replaced by memory cells, which allow the network to maintain longer temporal sequences. Thus, they improve on feed-forward neural networks by incorporating temporal information while avoiding the vanishing gradient problem of recurrent neural networks. Bi-directional LSTMs or BLSTMs are a further extension that model both past and future information. Figure 7 shows a BLSTM network in which hidden layers are connected both forwards and backwards. Features from individual modalities are concatenated in the input layer in LSTMs or BLSTMs. However, we do not consider this to be feature-level fusion as the hidden layers maintain a sophisticated internal model of the incoming data and the networks internal context.

LSTMs and BLSTMs have been successful with modalities such as speech where longer context

can provide significant discriminative power. For example, Eyben et al. [2010] fused acoustic and linguistic features in a BLSTM to classify affect in an evaluation-activation space, finding that it outperformed a basic recurrent neural network. Ringeval et al. [2015a] fused video, audio, and physiology and showed advantages of LSTMs and BLSTMs compared to feed-forward neural networks (this study is discussed in more detail below).

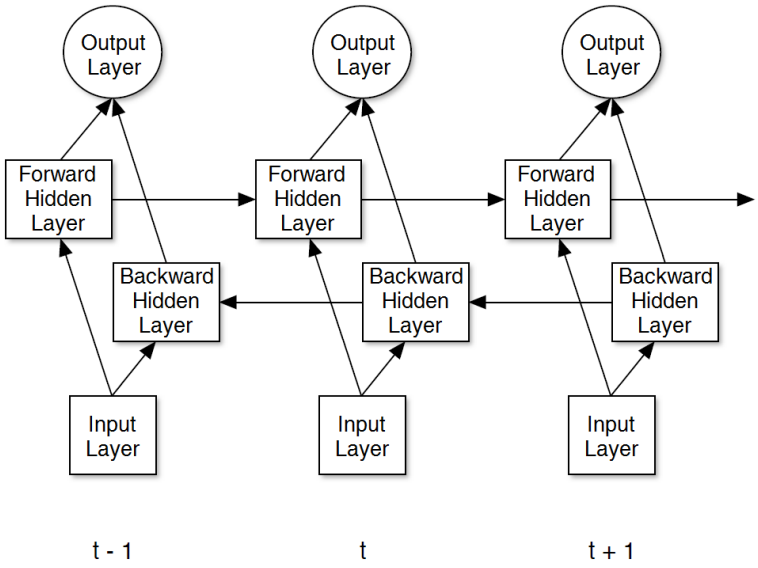

Figure 7. BLSTM with memory cells in each hidden layer

More recently, deep neural networks are being increasingly used for modality fusion in MMAD systems [Le Cun et al. 2015]. Deep networks contain multiple hidden layers and are capable of learning feature representations from raw data. For example, Kahou et al. [2013] used deep neural networks to classify affect from several modalities including video and audio. They first trained separate deep networks for each modality, then fused the networks together by weighting each network in a final prediction.

The extremely large amounts of data required for deep learning are difficult to acquire in affect detection applications. However, a two-step approach can be employed to decrease the need for large affect databases (although this is more common for video rather than other modalities). First, deep networks that have been trained for more general classification tasks (e.g., object recognition) are obtained (presumably one for each modality). Second, affect detectors are developed by combining the last few hidden layers from each deep network into a new final layer and training that final layer using affect databases (example network in Figure 8). This method utilizes the sparse feature representations that have been learned by the deep networks in their deeper hidden layers without requiring prohibitively large affect databases, and can be considered a form of transfer learning. For example, Ng et al. [2015] found a 16% improvement by fine-tuning an object recognition deep network using multiple affect databases versus training on only one affect database.

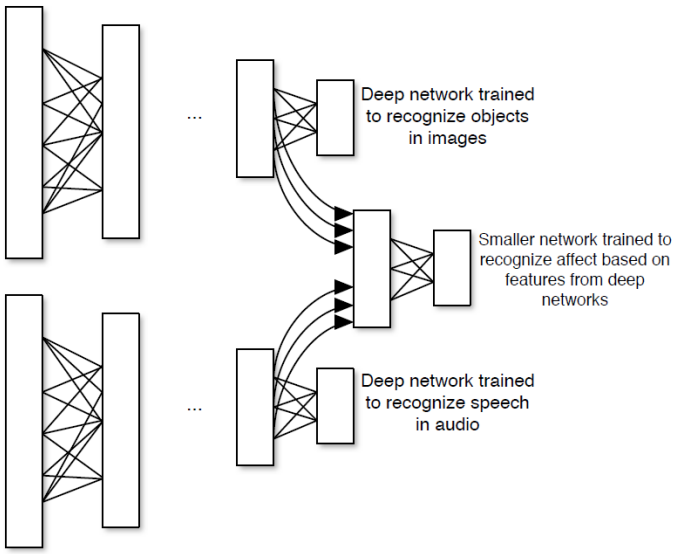

Figure 8. Fusion of deep neural networks by re-training final layers from networks representing each modality

X.1 Walk-throughs of sample multisensor-multimodal affect detection systems

We present three walk-throughs to serve as concrete renditions of MMAD systems. The walk-throughs were selected to emphasize the wide variability of research in the area and to highlight the various challenges and design decision facing MMAD systems.

Walk-through 1 – Feature-level fusion for detection of basic emotions

Our first walk-through was concerned with detection of emotions elicited through an affect elicitation procedure. Janssen et al. [2013] compared automatic detection vs. human perception of three basic emotions (happy, sad, angry), relaxed, and neutral, which were induced via an autobiographical recall procedure [Baker and Guttfreund 1993]. According to this procedure, 17 *stimulus subjects* were asked to write about two events in their life associated with experiences of these emotions. They were then asked to recall a subset of those events in a way that made them relive the emotions experienced. They then verbally described each event (in Dutch) in 2-3 minute trials. Audio, video, and physiological signals (electrodermal activity, skin temperature, respiration, and electrocardiography) were recorded while the stimulus subjects recalled and described the events. Each recording was associated with the label of the corresponding emotion being recalled, which was taken to be the "ground truth."

The authors extracted a variety of features from the signals. Facial features included movement of automatically tracked facial landmarks around the mouth and the eyes, as well as head position. Standard acoustic-prosodic features (e.g., fundamental frequency (pitch), energy, jitter, shimmer, formants) were extracted from the speech signal. Example physiological features included respiration rate, interbeat intervals, mean skin temperature, and number of skin conductance responses. A support vector machine classifier was trained to discriminate among the elicited emotions (five-way classification) using features from the individual modalities as well from

feature-level modality fusion and best-first search (see Figure 9). The multimodal model obtained a classification accuracy of 82%, which was greater than the individual modalities: 39% for audio, 59% for video, and 76% for physiology.

The authors compared computer vs. human affect detection accuracy. This was done by asking a set of human judges to classify the elicited emotions based on various stimuli combinations (audio-only, video-only, audio-video). Both U.S. and Dutch judges were used, but we only report results from the Dutch judges since they match the stimulus subjects. The Dutch judges were the most accurate (63%) when provided with audio (which was also in Dutch), compared to video (36%), and combined audio-video (48%). However, their accuracy was considerably lower than the automated detector (82%), although this result should be interpreted with caution as the testing protocols may have been biased in favor of the computer as strict person-level independence between training and testing sets was not enforced. Nevertheless, this remains one of the few studies that has contrasted human- vs. machine- classification on a multimodal dataset.

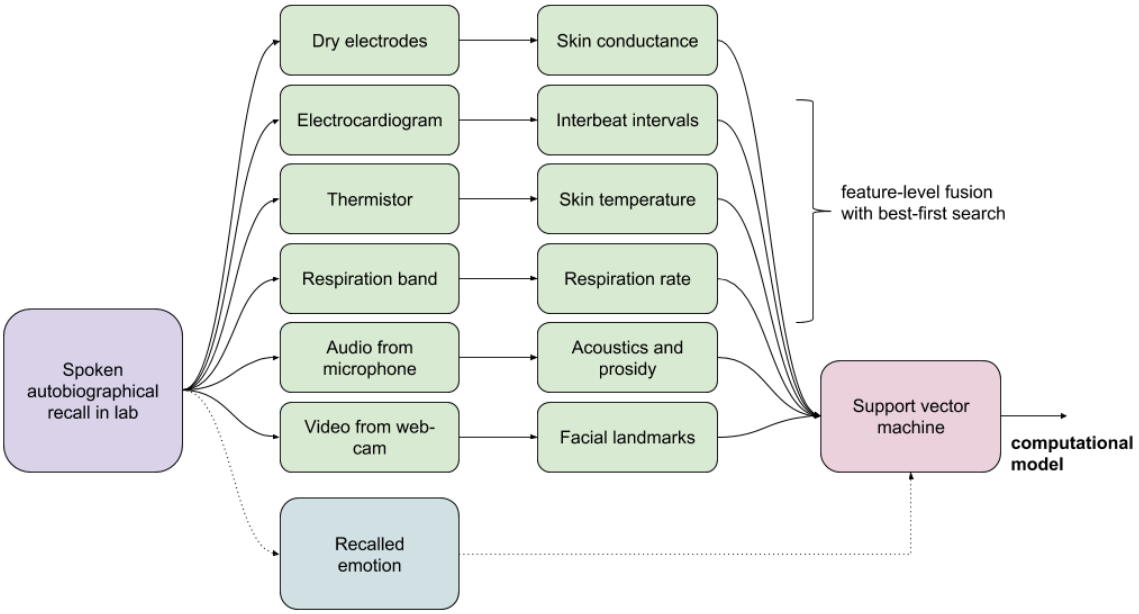

Figure 9. Schematic for walk-through 1

Walk-through 2 – Decision-level fusion for detection of learning-centered affective states

Our second walk-through focuses on multimodal affect detection in a computer-enabled classroom [Bosch et al. 2015b]. The researchers collected training data from 137 (8th and 9th grade) U.S. students who learned from a conceptual physics educational game called Physics Playground [Shute et al. 2013]. Students played the game in two 55-minute sessions across two days. Trained observers performed live annotations of boredom, engaged concentration, confusion, frustration, and delight using the Baker-Rodrigo Observation Method Protocol (BROMP) [Ocumpaugh et al. 2012]. According to BROMP, the live annotations were based on observable behavior, including explicit actions towards the interface, interactions with peers and teachers, body movements,

gestures, and facial expressions. The observers had to achieve a kappa of 0.6 (inter-rater reliability) with an expert to be certified as a BROMP coder. Videos of students' faces and upper bodies and log files from the game were recorded and synchronized with the affect annotations.

The videos were processed using FACET – a computer-vision program [FACET 2014] which estimates the likelihood of 19 facial action units along with head pose and position. Body movement was also estimated from the videos using motion filtering algorithms [Kory et al. 2015]. Supervised learning methods were used to discriminate each affective state from the other states (e.g., boredom vs. confusion, frustration, engaged concentration, and delight) and were validated by randomly assigning students into training and testing sets across multiple iterations. The models yielded an average accuracy of 0.69 (measured with area under the receiver operating characteristic curve (AUROC or AUC), where a chance model could yield a value of 0.5). Follow-up validation analyses confirmed that the models generalized across multiple days (i.e., training on subset of students from day 1 testing on different students in day 2), class period, genders (i.e., training on males, testing on females and vice versa), and ethnicity as perceived by human coders [Bosch et al. 2016].

A limitation of video-based measures is that they are only applicable when the face can be detected in the video. This is not always the case outside of the lab, where there are occlusions, poor lighting, and other complicating factors. In fact, the face could only be detected about 65% of the time in this study. To address this, Bosch et al. [2015a] developed an additional computational model based on interaction/contextual features stored in the game log files (e.g., difficulty of the current game level, the student's actions, the feedback received, response times). The log-based models were less accurate (mean AUC of .57) than the video-based models (mean AUC of .67 after retraining), but could be applied in almost all of the cases. Separate logistic regression models were trained to adjudicate among the face- and log-based -models, essentially weighting their relative influence on the final outcome via stacking (see Figure 10). The resultant multimodal model was almost as accurate as the video-based model (mean AUC of .64 for multimodal vs .67 for face only), but was applicable almost all of the time (98% for multimodal vs. 65% for face only). These results are notable given the noisy nature of the real-world environment with students incessantly fidgeting, talking with one another, asking questions, and even occasionally using their cellphones. They also illustrate how a MMAD approach addressed a substantial missing data problem despite it not improving detection accuracy.

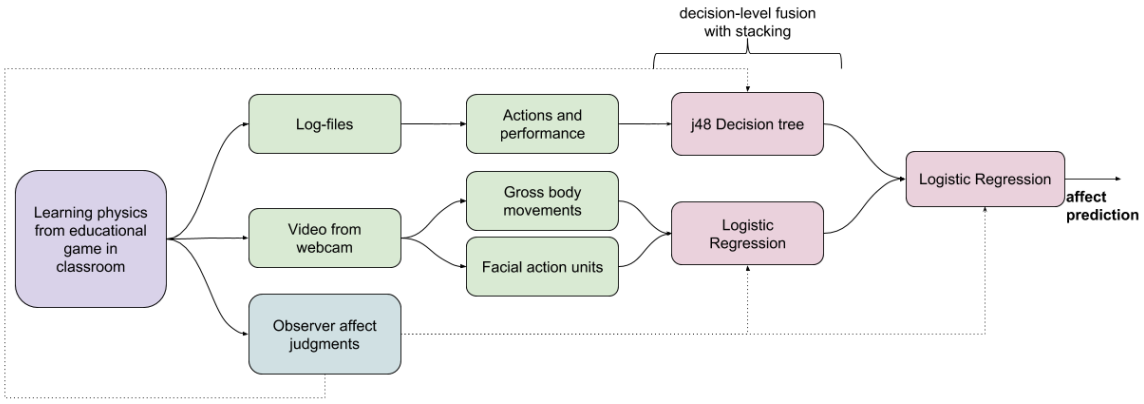

Figure 10. Schematic for walk-through 2

Walk-through 3 – Model-based fusion for modeling of affective dimensions

The previous two case studies focused on detecting discrete affective states with feature- or decision- level fusion. Our third walk-through used a neural network for modality fusion in the course of modeling time-continuous annotations of valence (unpleasant to pleasant) and arousal (sleepy to active) [Ringeval, Eyben, Kroupi, Yuce, Thiran, Ebrahimi, Lalanne and Schuller 2015a]. The authors recorded audio, facial video, electrocardiogram (ECG), and electro-dermal activity (EDA) as dyads completed a "winter survival" collaborative task. A total of 46 participants completed the task, of whom 34 provided permission for their data to be used. Data from a further 7 participants had recording errors, yielding a final data set of 27 participants. Six observers annotated the first five minutes of each participant's data by providing time-continuous ratings of valence and arousal. The recordings and annotations are distributed as part of the RECOLA dataset [Ringeval et al. 2013], which has been used in recent MMAD challenges [Ringeval et al. 2015b].

A variety of features were extracted from each of the modalities (audio, video, ECG, and EDA). Audio features captured spectral, prosodic, and voice quality metrics. Video features included 15 automatically extracted facial action units (AUs) and head pose. ECG features primarily consisted of heart rate, heart rate variability, and spectral features. EDA features mainly emphasized changes in skin conductance. LSTM and BLSTM networks (as discussed above) were trained to estimate continuous valence and arousal annotations by fusing features from the various modalities (Figure 11). The networks were validated in a person-independent fashion. The concordance correlation r_c (combining Pearson's r and mean squared error) was used to measure model accuracy.

The authors performed several experiments, including both early and late fusion and various combinations of modalities; here we focus on each feature (from any modality or combination) being an input node in the network. The best model-level fusion achieved a r_c of .769 for arousal and a r_c of .492 for valence. These best results were obtained using a combination of audio and video features. Further, when compared to standard feed-forward neural networks, the BLSTM models were more accurate across shorter windows of time (2-3 secs) but accuracy was equitable across longer windows (4-5 secs). Finally, when compared to individual modalities, there was a multimodal advantage for valence (r_c = .492 vs. .431), but not for arousal (r_c = .769 vs. .788), once

again highlighting selective conditions where MMAD led to improvements over UMAD.

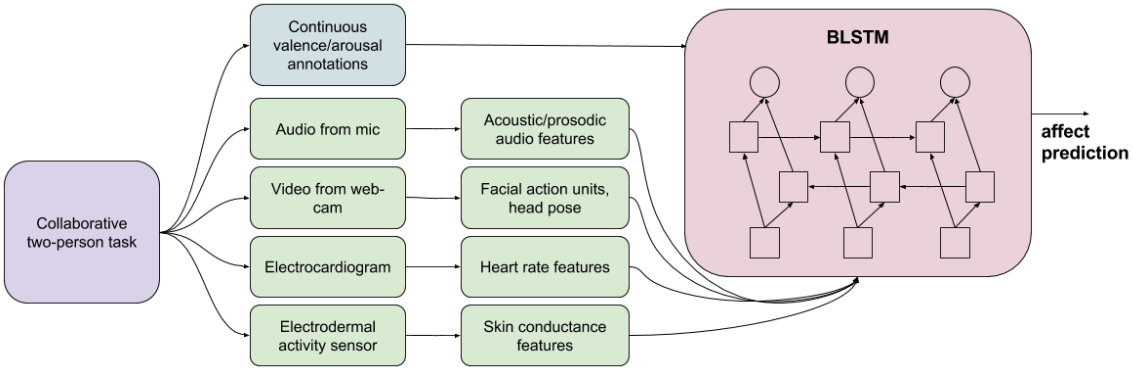

Figure 11. Schematic for walk-through 3

X.1 General trends and state of the art in multisensor-multimodal affect detection

D'Mello and Kory [2015] recently performed a review and meta-analysis of 90 MMAD systems. We highlight some of their key findings, both in terms of trends in MMAD system design as well as classification accuracy of MMAD vs. UMAD. Table 2 lists a subset (about 1/3) of the more recent studies (2011 to 2013) reviewed in D'Mello and Kory [2015] along with a few more recent studies (2014-) published since their review.

Table 2. Selective sample of recent MMAD systems in the D'Mello and Kory [2015] review (2011 to 2013), further extended to include more recent systems (2014 and 2015)

Reference	Modalities	Fusion
[Chanel et al. 2011]	EEG + Physiology	Decision
[Datcu and Rothkrantz 2011]	Face + Voice	Feature
[Jiang, Cui, Zhang, Fan, Ganzalez and Sahli 2011]	Face + Voice	Model
[Lingenfelser et al. 2011]	Face + Voice	Decision
[Nicolaou et al. 2011]	Face + Voice + Body	Model
[Schuller 2011]	Voice + Text	Feature
[Vu et al. 2011]	Voice + Body	Decision
[Wagner et al. 2011]	Face + Voice + Body	Decision
[Walter et al. 2011]	Voice + Physiology	Decision
[Wu and Liang 2011]	Voice + Text	Decision

Reference	Modalities	Fusion
[Hussain et al. 2012]	Face + Physiology	Decision
[Koelstra et al. 2012]	EEG + Physiology + Content	Decision
[Lin, Wu and Wei 2012]	Face + Voice	Model
[Lu and Jia 2012]	Face + Voice	Model
[Metallinou et al. 2012]	Face + Voice	Model
[Monkaresi et al. 2012]	Face + Physiology	Feature
[Park et al. 2012]	Face + Voice	Decision
[Rozgic et al. 2012]	Face + Voice + Text	Feature
[Savran et al. 2012]	Face + Voice + Text	Model
[Soleymani et al. 2012]	EEG + Gaze	Decision
[Baltrušaitis et al. 2013]	Face + Voice	Model
[Dobrišek et al. 2013]	Face + Voice	Decision
[Glodek et al. 2013]	Face + Voice	Decision
[Hommel et al. 2013]	Face + Voice	Decision
[Krell et al. 2013]	Face + Voice	Decision
[Rosas et al. 2013]	Face + Voice + Text	Feature
[Rosas, Mihalcea and Morency 2013]	Face + Voice + Text	Feature
[Wang et al. 2013]	EEG + Content	Feature
[Wöllmer et al. 2013a]	Face + Voice	Model
[Wöllmer et al. 2013b]	Face + Voice + Text	Hybrid
[Williamson et al. 2014]	Face + Voice	Decision
[Grafsgaard et al. 2014]	Face + Posture + Interaction	Feature
[Soleymani et al. 2014]	Face + EEG	Model
[Bosch, Chen, Baker, Shute and D'Mello 2015a]	Face + Interaction	Decision
[Zhou et al. 2015]	Face + Interaction + Content	Feature

Reference	Modalities	Fusion
[Barros et al. 2015]	Face + Body	Model
[Monkaresi et al. 2017]	Face + Remote Physiology	Decision

Note. Physiology refers to one or more peripheral physiological channels such as electrodermal activity, heart rate variability, etc.

Trends in MMAD systems

D'Mello and Kory [2015] coded each MMAD system across a number of dimensions, such as whether the training data consisted of acted, induced, or naturalistic affective expressions, the specific modality combinations used, the most successful fusion method, and so on. Below are some of the highlights of MMAD as of 2013.

- MMAD systems were trained on small samples. The studies had on average of 21 participants and 97% of the studies had fewer than 50 participants.
- Training data for about half the studies were obtained by actors portraying affective expressions. Affective states were induced in 28% of the studies using validated elicitation methods [Coan and Allen 2007]. Very few studies (20% of studies) used naturalistic affective states (i.e., affective states that spontaneously arise as part of an interaction).
- In terms of MMAD, bimodal systems were far more common (87%) than trimodal systems (13%).
- The face and voice (paralinguistics) were the two most frequent modalities, each occurring in over 75% of the studies. By comparison, peripheral physiology was only used in 11% of the systems and other modalities (e.g., eye tracking) were much rarer.
- About a 1/3 of the studies (37%) focused on detecting the basic emotions of anger, fear, happiness, sadness, disgust, and surprise [Ekman 1992] or core affective dimensions of valence and arousal (28%). Very few studies focused on detecting additional affect dimensions, such as dominance or certainty [Fontaine, Scherer, Roesch and Ellsworth 2007] or nonbasic affective states like confusion and curiosity [D'Mello and Calvo 2013].
- Feature-level (39%) and decision-level (35%) fusion were much more common than hybrid (6%) and model-level fusion (20%)
- A vast majority of studies employed instance-level validation (62%), where *different* instances from the *same* person were in both training and test sets, essentially limiting generalizability to new individuals.

Accuracy of MMAD systems

How accurate are MMAD systems compared to their unimodal affect detection (UMAD) counterparts? D'Mello and Kory [2015] addressed this question by computing the percent improvement in classification accuracy of each MMAD system compared to the best UMAD system (called MM1 effects). They also investigated factors that moderated MM1 effects. Their key findings indicated that:

- On average, MMAD yielded a 10% improvement in affect detection accuracy over the best UMAD counterpart.

- There were negative or negligible (<= 1%) MM1 effects for 14.4% of the studies, about 50% yielded small 1-5% or medium-sized (5-10%) effects, while the remaining 35% yielded impressively large effects (> 10%).
- The median MM1 effect of 7% might be a more accurate estimate given the spread of the distribution.
- There was a very robust correlation (Pearson's r = .87) between best UMAD and MMAD accuracies, suggesting a high degree of redundancy; see Vinciarelli and Esposito [2017].
- The mean MM1 effect for detectors trained on naturalistic data (4.6%) was three times lower compared to detectors trained on acted data (12.7%) and about half compared to detectors trained on experimentally induced affective states (8.2%).
- Model-based fusion methods resulted in a roughly twice the mean MM1 effect (15.3%) compared to feature-level (7.7%) and decision-level (6.7%) fusion. However, this result should be taken with a modicum of caution because it involves between- study comparisons where additional factors could have varied.

Importantly, the authors were able to predict MMAD accuracy from best UMAD accuracy using data type (1 for acted data; 0 for induced or naturalistic data) and fusion method (1 for model-level fusion; 0 for feature- or decision- level fusion). The regression model shown (using standardized coefficients) below explained an impressive 83.3% of the variance based on 10-fold study-level cross-validation.

MMAD accuracy = .900 × Best UMAD accuracy + .273 × Data Type Acted [1 or 0] + .312 × Model Level Fusion [1 or 0] -.253

MMAD Systems from the 2015 Audio-Video Emotion Recognition Challenge (AV+EC 2015)

The Audio-Video Emotion Recognition Challenge (AVEC) series is an annual affect detection competition that was first organized as part of the 2011 Affective Computing and Intelligent Interaction (ACII) conference series [Schuller et al. 2011]. The earlier challenges emphasized audio-visual detection of time-continuous annotations of affective dimensions [Schuller et al. 2012] based on data from the SEMAINE corpus [McKeown et al. 2012], which was designed to collect naturalistic data of humans interacting with artificial agents. The most recent challenge (at the time of writing) was the Audio-Visual+ Emotion recognition Challenge and workshop (AV+EC 2015), where the goal was to model time-continuous annotations of valence and arousal from audio, video, and physiology (electrocardiogram and electrodermal activity) signals collected as part of the RECOLA data set [Ringeval, Sonderegger, Sauer and Lalanne 2013] (see walk-through 3 above).

Table 3 presents the seven MMAD systems featured in the AV+EC 2015 challenge. Two systems adopted a UMAD approach and are not included here. We note the popularity of model-based fusion techniques, especially those using LSTMs and their variants, although feature- and decision-level fusion methods still feature quite prominently. The best result was obtained by He et al. [2015], who adopted a deep (i.e., multilayer) BLSTM for modality fusion. They achieved a concordance correlation (r_c - see walk-through 3) of .747 for arousal and .609 for valence, both reflecting substantial improvements over the challenge baselines (r_c = .444 for arousal and .382 for valence).

Table 3. MMAD systems featured in the AV+EC 2015 challenge

Reference	Fusion Method
[Cardinal et al. 2015]	Feature, Decision (random forest, linear regression)
[Milchevski et al. 2015]	Feature, Decision (linear regression)
[Huang et al. 2015]	Feature, Decision (linear regression), Hybrid
[Chen and Jin 2015]	Model (BLSTM)
[Chao et al. 2015]	Model (LSTM)
[He, Jiang, Yang, Pei, Wu and Sahli 2015]	Model (Deep BLSTM)
[Kächele et al. 2015]	Feature, Decision (averaging), Model (multilayer perceptron)

X.1 Discussion

At the time of this writing, affective computing is nearing its 20 year birthdate [Picard 1997] (see Picard [2010] for a brief history of the field). In D'Mello and Kory [2015], we summarized the state of the field of affect detection in 2003 as:

> "the use of basic signal processing and machine learning techniques, independently applied to still frames (but occasionally to sequences) of facial or vocal data, to detect exaggerated context-free expressions of a few basic affective states that are acted by a small number of individuals with no emphasis on generalizability."

It is clear as much progress has been made over the next 10 years as noted by our summary of the field as of 2013. The italicized items highlight key changes from 2003 to 2013. Most notable is the shift in emphasis from facial *or* vocal signals to facial *and* vocal signals, suggesting that we are finally in the age of MMAD, despite sustained progress in UMAD.

> "the use of basic *and advanced* signal processing and machine learning techniques, independently *and jointly* applied to *sequences* of *primarily* facial *and* vocal data, to detect exaggerated *and naturalistic* context-free *and context-sensitive* expressions of a *modest* number of basic affective states and *simple dimensions* that are acted *or experienced* by a *modest* number of individuals with *some* emphasis on generalizability."

What would be a prospective summary of the field a decade from now - say in 2027? We anticipate progress in data collection methods (sensors used, modalities considered, data collection contexts, size of data sets), the computational methods (signal processing, machine learning, fusion techniques), and the affective phenomenon itself (affective states modeled, affect representations, how "ground truth" is established).

But what about the metrics of success? The metrics we utilize embody what we (as a community) *value* in affect detection systems. It is fair to say that detection (or prediction) accuracy on unseen

data is the key metric of success in the field (e.g., the AV+EC challenge selects winners based on prediction accuracy on a held-out test set). Does accuracy, then, embody our values?

If so, then one must ask "accurate for what purpose and in what context?" Is a highly accurate system trained on a handful of participants in a lab setting of more value than a less accurate one trained on noisy data, but from thousands of individuals in the wild? Similarly, is a highly accurate system that cannot function in the presence of missing data of more value than its less accurate counterpart that is robust to data loss? If accuracy is not the only metric that embodies our values, then what might be some alternative metrics?

The answer might lie into the very nature of affect itself. Recall that affect is a construct, not a physical entity. It cannot be precisely defined or directly measured, but only approximated. This level of imprecision might be discomforting to some who might rightly ask: "how can we measure what we cannot even define?" This question has plagued researchers in the psychological sciences for several decades, who have proposed a host of metrics, each based on different criterion of success. These include different forms of reliability, convergent validity (closely related to accuracy), discriminant validity, ecological validity (related to generalizability), predictive validity, criterion validity, and so on [Rosenthal and Rosnow 1984].

Herein lies the rub. Many of these criteria are in a state of tension. A system (or measure) that achieves impressive gains along one criterion likely does so at the expense of another. Want a highly accurate (but not very generalizable) system? Just lock a few participants in the lab and ask them to act out a couple of emotions. Want a generalizable (but not very accurate) system? Try to capture affective expressions as people go about their daily routines in the world. By considering a range of metrics, we are forced to identify the inherent weaknesses in our systems and confront out assumptions about the nature of affect and "affective ground truth." Thus, in addition to anticipated advances in theoretical sophistication, data sources, and computational techniques, we advocate for an equitable advance in the science of validation over the next decade of multisensor-multimodal affect detection research. Only then will we have a chance of developing affect detection systems that will break through the confines of the lab and live up to their fullest potential in the real world.

Acknowledgments

This research was supported by the National Science Foundation (NSF) (DRL 1235958 and IIS 1523091). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

References

E. André. Real-time Sensing of Affect and Social Signals in a Multimodal Context. In *Handbook of Multimodal-Multisensor Interfaces*, 2017

R.C. Baker and D.O. Guttfreund. The effects of written autobiographical recollection induction procedures on mood. *Journal of Clinical Psychology 49*, 563-568, 1993

T. Baltrušaitis, N. Banda and P. Robinson. Dimensional Affect Recognition using Continuous Conditional Random Fields. In *Proceedings of the International Conference on Multimedia and Expo (Workshop on Affective Analysis in Multimedia)*, 2013

L. Barrett. Are emotions natural kinds? *Perspectives on Psychological Science 1*, 28-58, 2006

L. Barrett, B. Mesquita, K. Ochsner and J. Gross. The experience of emotion. *Annual Review of*

Psychology 58, 373-403, 2007
L.F. Barrett. The conceptual act theory: A précis. *Emotion Review 6*, 292-297, 2014
P. Barros, D. Jirak, C. Weber and S. Wermter. Multimodal emotional state recognition using sequence-dependent deep hierarchical features. *Neural Networks 72*, 140-151, 2015
N. Bosch, H. Chen, R. Baker, V. Shute and S.K. D'Mello. Accuracy vs. Availability Heuristic in Multimodal Affect Detection in the Wild. In *Proceedings of the 17th ACM International Conference on Multimodal Interaction (ICMI 2015)* ACM, New York, NY, 2015a
N. Bosch, S.K. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura and L. Wang. Automatic Detection of Learning-Centered Affective States in the Wild. In *Proceedings of the 2015 International Conference on Intelligent User Interfaces (IUI 2015)* ACM, New York, NY, 379-388, 2015b
N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh and V. Shute. Using video to automatically detect learner affect in computer-enabled classrooms. *ACM Transactions on Interactive Intelligent Systems 6*, 17.11-17.31, 2016
M.M. Bradley and P.J. Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry 25*, 49-59, 1994
R. Calvo, S.K. D'Mello, J. Gratch and A. Kappas. The Oxford Handbook of Affective Computing Oxford University Press, New York, NY, 2015
R.A. Calvo and S.K. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing 1*, 18-37, 2010
L. Camras and J. Shutter. Emotional facial expressions in infancy. *Emotion Review 2(2)*, 120-129, 2010
P. Cardinal, N. Dehak, A.L. Koerich, J. Alam and P. Boucher. ETS system for AV+EC 2015 challenge. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 17-23, 2015
G. Chanel, C. Rebetez, M. Bétrancourt and T. Pun. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 41*, 1052-1063, 2011
L. Chao, J. Tao, M. Yang, Y. Li and Z. Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 65-72, 2015
D. Chen, D. Jiang, I. Ravyse and H. Sahli. Audio-visual emotion recognition based on a DBN model with constrained asynchrony. . In *Proceedings of the Fifth International Conference on Image and Graphics (ICIG 09)* IEEE, Washington, DC, 912–916, 2009
S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 49-56, 2015
J. Coan and J. Allen. Handbook of emotion elicitation and assessment Oxford University Press, New York, 2007
J.A. Coan. Emergent ghosts of the emotion machine. *Emotion Review 2*, 274-285, 2010
C. Conati and H. Maclaren. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction 19*, 267-303, 2009
R. Cowie, G. McKeown and E. Douglas-Cowie. Tracing emotion: an overview. *International Journal of Synthetic Emotions (IJSE) 3*, 1-17, 2012

S. D'Mello and R. Calvo. Beyond the Basic Emotions: What Should Affective Computing Compute? In *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2013)*, S. Brewster, S. Bødker and W. Mackay Eds. ACM, New York, NY, 2013

S.K. D'Mello. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing 7*, 136-149, 2016

S.K. D'Mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys 47*, 43:41-43:46, 2015

S. D'Mello and A. Graesser. The half-life of cognitive-affective states during complex learning. *Cognition & Emotion 25*, 1299-1308, 2011

S.K. D'Mello, N. Dowell and A.C. Graesser. Unimodal and multimodal human perception of naturalistic non-basic affective states during Human-Computer interactions. *IEEE Transactions on Affective Computing 4*, 452 - 465, 2013

A. Damasio. *Looking for Spinoza: Joy, sorrow, and the feeling brain*. Harcourt Inc., 2003

C. Darwin. *The expression of the emotions in man and animals*. John Murray, London, 1872

D. Datcu and L. Rothkrantz. Emotion recognition using bimodal data fusion. In *Proceedings of the 12th International Conference on Computer Systems and Technologies* ACM, New York, NY, 122-128, 2011

S. Dobrišek, R. Gajšek, F. Mihelič, N. Pavešić and V. Štruc. Towards Efficient Multi-Modal Emotion Recognition. *International Journal of Advanced Robotic Systems 10*, 1-10, 2013

P. Ekman. An argument for basic emotions. *Cognition & Emotion 6*, 169-200, 1992

P. Ekman. Strong Evidence for Universals in Facial Expressions - a Reply to Russells Mistaken Critique. *Psychological Bulletin 115*, 268-287, 1994

H. Elfenbein and N. Ambady. Is there an ingroup advantage in emotion recognition? *Psychological Bulletin 128*, 243-249, 2002a

H. Elfenbein and N. Ambady. On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin 128*, 203-235, 2002b

F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie and R. Cowie. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces 3*, 7-19, 2010

FACET. Facial Expression Recognition Software Emotient, Boston, MA, 2014

J. Fontaine, K. Scherer, E. Roesch and P. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science 18*, 2007

A.J. Fridlund, P. Ekman and H. Oster. Facial expressions of emotion. In *Nonverbal behavior and communication*, A.W. Siegman and S. Feldstein Eds. Erlbaum, Hillsdale, NJ, 143–223, 1987

J.M. Girard, J.F. Cohn, L.A. Jeni, M.A. Sayette and F. De la Torre. Spontaneous facial expression in unscripted social interactions can be measured automatically. *Behavior Research Methods 47*, 1136-1147, 2015

M. Glodek, S. Reuter, M. Schels, K. Dietmayer and F. Schwenker. Kalman Filter Based Classifier Fusion for Affective State Recognition. In *Proceedings of the 11th International Workshop on Multiple Classifier Systems*, Z.-H. Zhou, F. Roli and J. Kittler Eds. Springer, Berlin Heidelberg, 85-94, 2013

J.F. Grafsgaard, J.B. Wiggins, K.E. Boyer, E.N. Wiebe and J.C. Lester. Predicting learning and affect from multimodal data streams in task-oriented tutorial dialogue. In *Proceedings of*

the 7th International Conference on Educational Data Mining, J. Stamper, Z. Pardos, M. Mavrikis and B.M. McLaren Eds. International Educational Data Mining Society, 122-129, 2014

J.J. Gross and L.F. Barrett. Emotion generation and emotion regulation: One or two depends on your point of view. *Emotion Review 3*, 8-16, 2011

L. He, D. Jiang, L. Yang, E. Pei, P. Wu and H. Sahli. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 73-80, 2015

S.J. Heine, D.R. Lehman, K. Peng and J. Greenholtz. What's wrong with cross-cultural comparisons of subjective Likert scales?: The reference-group effect. *Journal of Personality and Social Psychology 82*, 903-918, 2002

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation 9*, 1735-1780, 1997

S. Hommel, A. Rabie and U. Handmann. Attention and Emotion Based Adaption of Dialog Systems. In *Intelligent Systems: Models and Applications*, E. Pap Ed. Springer Verlag, Berlin Heidelberg, 215-235, 2013

Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu and J. Epps. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 41-48, 2015

M. Hussain, H. Monkaresi and R. Calvo. Combining Classifiers in Multimodal Affect Detection. In *Proceedings of the Australasian Data Mining Conference*, 2012

C. Izard. Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin 115*, 1994

C. Izard. The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review 2*, 363-370, 2010

C.E. Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science 2*, 260-280, 2007

W. James. What is an emotion? *Mind 9*, 188-205, 1884

J.H. Janssen, P. Tacken, J. de Vries, E.L. van den Broek, J.H. Westerink, P. Haselager and W.A. IJsselsteijn. Machines outperform laypersons in recognizing emotions elicited by autobiographical recollection. *Human–Computer Interaction 28*, 479-517, 2013

D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez and H. Sahli. Audio visual emotion recognition based on triple-stream dynamic bayesian network models. In *Proceedings of the Fourth International Conference on Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, S. B and J. Martin Eds. Springer-Verlag, Berlin Heidelberg, 609-618, 2011

M. Kächele, P. Thiam, G. Palm, F. Schwenker and M. Schels. Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 9-16, 2015

S.E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio and R.C. Ferrari. Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM International Conference on Multimodal Interaction* ACM, New York, 543-550, 2013

S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt and I. Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing 3*, 18-31, 2012

J. Kory and S.K. D'Mello. Affect elicitation for affective computing. In *The Oxford Handbook of Affective Computing*, R. Calvo, S. D'Mello, J. Gratch and A. Kappas Eds. Oxford University Press, New York, NY., 371-383, 2015

J. Kory, S.K. D'Mello and A. Olney. Motion Tracker: Camera-based Monitoring of Bodily Movements using Motion Silhouettes. *Plos One 10*, 10.1371/journal.pone.0130293, 2015

G. Krell, M. Glodek, A. Panning, I. Siegert, B. Michaelis, A. Wendemuth and F. Schwenker. Fusion of Fragmentary Classifier Decisions for Affective State Recognition. In *Proceedings of the The 1st International Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction*, F. Schwenker, S. Scherer and L.-P. Morency Eds. Springer-Verlag, Berlin Heidelberg, 116-130, 2013

J.A. Krosnick. Survey research. *Annual Review of Psychology 50*, 537-567, 1999

Y. Le Cun, Y. Bengio and G.E. Hinton. Deep learning. *Nature 521*, 436-444, 2015

H.C. Lench, S.W. Bench and S.A. Flores. Searching for evidence, not a war: Reply to Lindquist, Siegel, Quigley, and Barrett (2013). *Psychological Bulletin 113*, 264-268, 2013

J.S. Lerner and D. Keltner. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion 14*, 473-493, 2000

M.D. Lewis. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behavioral and Brain Sciences 28*, 169-245, 2005

X. Li and Q. Ji. Active affective state detection and user assistance with dynamic bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 35*, 93–105, 2005

J. Lin, C. Wu and W. Wei. Error Weighted Semi-Coupled Hidden Markov Model for Audio-Visual Emotion Recognition. *IEEE Transactions on Multimedia 14*, 142 -156, 2012

K.A. Lindquist, A.B. Satpute, T.D. Wager, J. Weber and L.F. Barrett. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. *Cerebral Cortex 26*, 1910-1922, 2016

K.A. Lindquist, E.H. Siegel, K.S. Quigley and L.F. Barrett. The Hundred-Year Emotion War: Are Emotions Natural Kinds or Psychological Constructions? Comment on Lench, Flores, and Bench (2011). *Psychological Bulletin 139*, 264-268, 2013

K.A. Lindquist, T. Wager, D., H. Kober, E. Bliss-Moreau and L.F. Barrett. The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences 173*, 1-86, 2011

F. Lingenfelser, J. Wagner and E. André. A systematic discussion of fusion techniques for multi-modal affect recognition tasks. In *Proceedings of the 13th International Conference on Multimodal Interfaces* ACM, New York, NY, 19-26, 2011

M. Liu, R. Wang, S. Li, S. Shan, Z. Huang and X. Chen. Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th ACM International Conference on Multimodal Interaction* ACM, New York, 494-501, 2014

G. Loewenstein and J.S. Lerner. The role of affect in decision making. *Handbook of affective science 619*, 3, 2003

K. Lu and Y. Jia. Audio-visual emotion recognition with boosted coupled HMM. In *Proceedings of the 21st International Conference on Pattern Recognition* IEEE, Washington, DC,

1148-1151, 2012
J.-C. Martin, C. Clavel, M. Courgeon, M. Ammi, M.-A. Amorim, Y. Tsalamlal and Y. Gaffary. How Do Users Perceive Multimodal Expressions of Affects? In *Handbook of Multimodal-Multisensor Interfaces*, 2017
G. McKeown, M. Valstar, R. Cowie, M. Pantic and M. Schroder. The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing 3*, 5-17, 2012
M. Mehu and K. Scherer. A psycho-ethological approach to social signal processing. *Cognitive Processing 13*, 397-414, 2012
B. Mesquita and M. Boiger. Emotions in context: A sociodynamic model of emotions. *Emotion Review 6*, 298-302, 2014
A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller and S. Narayanan. Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification. *IEEE Transactions on Affective Computing 3*, 184-198, 2012
A. Milchevski, A. Rozza and D. Taskovski. Multimodal affective analysis combining regularized linear regression and boosted regression trees. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 33-39, 2015
H. Monkaresi, N. Bosch, R.A. Calvo and S.K. D'Mello. Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing 8*, 15-28, 2017
H. Monkaresi, M.S. Hussain and R. Calvo. Classification of affects using head movement, skin color features and physiological signals. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics* IEEE, Washington, DC, 2664-2669, 2012
H.-W. Ng, V.D. Nguyen, V. Vonikakis and S. Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI 2015)* ACM, New York, 443-449, 2015
M. Nicolaou, H. Gunes and M. Pantic. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence& Arousal Space. *IEEE Transactions on Affective Computing 2*, 92-105, 2011
J. Ocumpaugh, R.S. Baker and M.M.T. Rodrigo. Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0 Worcester Polytechnic Institute, Teachers College Columbia University, & Ateneo de Manila University, 2012
J. Ocumpaugh, R.S. Baker and M.M.T. Rodrigo. Baker Rodrigo Ocumpaugh Monitoring Protocol (BROMP) 2.0 Technical and Training Manual Teachers College, Columbia University, and Ateneo Laboratory for the Learning Sciences, New York, NY and Manila, Philippines, 2015
J. Park, G. Jang and Y. Seo. Music-aided affective interaction between human and service robot. *EURASIP Journal on Audio, Speech, and Music Processing 2012*, 1-13, 2012
B. Parkinson, A.H. Fischer and A.S. Manstead. *Emotion in social relations: Cultural, group, and interpersonal processes*. Psychology Press, 2004
R. Picard. *Affective Computing*. MIT Press, Cambridge, Mass, 1997
R. Picard. Affective Computing: From Laughter to IEEE. *IEEE Transactions on Affective Computing 1*, 11-17, 2010
R.W. Picard, S. Fedor and Y. Ayzenberg. Multiple arousal theory and daily-life electrodermal activity asymmetry. *Emotion Review*, 2015

P.M. Podsakoff, S.B. MacKenzie, J.Y. Lee and N.P. Podsakoff. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology 88*, 879-903, 2003

F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters 66*, 22-30, 2015a

F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie and M. Pantic. AV+ EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, 3-8, 2015b

F. Ringeval, A. Sonderegger, J. Sauer and D. Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proceedings of the 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE) in conjunction with the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition* IEEE, Washington, DC, 2013

V. Rosas, R. Mihalcea and L. Morency. Multimodal Sentiment Analysis of Spanish Online Videos. *IEEE Intelligent Systems 28*, 38-45, 2013

I.J. Roseman. Emotional behaviors, emotivational goals, emotion strategies: Multiple levels of organization integrate variable and consistent responses. *Emotion Review 3*, 434-443, 2011

R. Rosenthal and R. Rosnow. *Essentials of behavioral research: Methods and data analysis*. McGraw-Hill, New York, 1984

V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar and R. Prasad. Ensemble of SVM trees for multimodal emotion recognition. In *Proceedings of the Signal & Information Processing Association Annual Summit and Conference* IEEE, Washington, DC, 1-4, 2012

W. Ruch. Will the real relationship between facial expression and affective experience please stand up: The case of exhilaration. *Cognition & Emotion 9*, 33-58, 1995

J. Russell. Core affect and the psychological construction of emotion. *Psychological Review 110*, 145-172, 2003

J.A. Russell, J.A. Bachorowski and J.M. Fernandez-Dols. Facial and vocal expressions of emotion. *Annual Review of Psychology 54*, 329-349, 2003

J.A. Russell, A. Weiss and G.A. Mendelsohn. Affect Grid - A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology 57*, 493-502, 1989

A. Savran, H. Cao, M. Shah, A. Nenkova and R. Verma. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction* ACM, New York, NY, 485-492, 2012

K.R. Scherer. The dynamic architecture of emotion: Evidence for the component process model. *Cognition & Emotion 23*, 1307-1351, 2009

B. Schuller. Recognizing Affect from Linguistic Information in 3D Continuous Space. *IEEE Transactions on Affective Computing 2*, 192-205, 2011

B. Schuller, M. Valster, R. Cowie and M. Pantic. AVEC 2011: Audio/Visual Emotion Challenge and Workshop. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011)*, S. D'Mello, A. Graesser, B. Schuller

and J.-C. Martin Eds. Springer, Berlin, 2011
B. Schuller, M. Valster, F. Eyben, R. Cowie and M. Pantic. AVEC 2012: The continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction* ACM, New York, NY, 449-456, 2012
V.J. Shute, M. Ventura and Y.J. Kim. Assessment and learning of qualitative physics in Newton's playground. *The Journal of Educational Research 106*, 423-430, 2013
M. Soleymani, S. Asghari-Esfeden, M. Pantic and Y. Fu. Continuous emotion detection using eeg signals and facial expressions. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* IEEE, Washington DC, 1-6, 2014
M. Soleymani, M. Pantic and T. Pun. Multi-Modal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing 3*, 211-223, 2012
S.S. Tomkins. *Affect Imagery Consciousness: Volume I, The Positive Affects*. Tavistock, London, 1962
J.L. Tracy. An evolutionary approach to understanding distinct emotions. *Emotion Review 6*, 308-312, 2014
A. Vinciarelli and A. Esposito. Multimodal Analysis of Social Signals. In *Handbook of Multimodal-Multisensor Interfaces*, 2017
H. Vu, Y. Yamazaki, F. Dong and K. Hirota. Emotion recognition based on human gesture and speech information using RT middleware. In *IEEE International Conference on Fuzzy Systems* IEEE, Washington, DC, 787-791, 2011
J. Wagner, E. Andre, F. Lingenfelser, J. Kim and T. Vogt. Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. *IEEE Transactions on Affective Computing 2*, 206-218, 2011
S. Walter, S. Scherer, M. Schels, M. Glodek, D. Hrabal, M. Schmidt, R. Böck, K. Limbrecht, H. Traue and F. Schwenker. Multimodal emotion classification in naturalistic user behavior. In *Proceedings of the International Conference on Human-Computer Interaction*, J. Jacko Ed. Springer, Berlin, 603-611, 2011
S. Wang, Y. Zhu, G. Wu and Q. Ji. Hybrid video emotional tagging using users' EEG and video content. *Multimedia Tools and Applications*, 1-27, 2013
J.R. Williamson, T.F. Quatieri, B.S. Helfer, G. Ciccarelli and D.D. Mehta. Vocal and Facial Biomarkers of Depression Based on Motor Incoordination and Timing. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* ACM, New York, NY, 65-72, 2014
M. Wöllmer, M. Kaiser, F. Eyben and B. Schuller. LSTM modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing 31*, 2013a
M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae and L. Morency. YouTube Movie Reviews:Sentiment Analysis in an Audiovisual Context. *IEEE Intelligent Systems 28*, 46-53, 2013b
C. Wu and W. Liang. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing 2*, 10-21, 2011
Z. Zeng, M. Pantic, G. Roisman and T. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Ieee Transactions on Pattern Analysis and Machine Intelligence 31*, 39-58, 2009
D. Zhou, J. Luo, V.M. Silenzio, Y. Zhou, J. Hu, G. Currier and H.A. Kautz. Tackling Mental

Health by Integrating Unobtrusive Multimodal Sensing. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-2015)* ACM, New York, 1401-1409, 2015

Supplementary Digital Materials

Table 4. Chapter Focus Questions

1. What do we mean when we say that affect is a multicomponential conceptual phenomenon?
2. Why is the affective experience-expression link weak and how is this related to loosely coupled uncoordinated affective responses?
3. Popular TV shows like “Lie to Me” assume that humans can be trained to be highly accurate emotion and deception detectors. Do you agree or disagree? Why?
4. Assume you want to develop a detector of surprise. What are three unique ways by which you could obtain affective ground truth to train your detector?
5. Assume you have three modalities: video, audio, and electrodermal activity. How would you combine them to achieve “hybrid fusion”?
6. Sketch four different model-level fusion designs that combine facial expressions, heart rate, eye movements, keystrokes, and user personality traits.
7. How would you estimate bimodal classification accuracy from corresponding unimodal classification accuracies without even building the multimodal model?
8. How would you go about building a multisensor-multimodal detector of *interest* while people read news articles on www.cnn.com? What about curiosity?
9. How would you build a robust multimodal-multisensor detector of confusion. Robust implies that the detector should operates even when some of the modalities do not provide any data.
10. The concluding section lists several metrics of success in addition to detection accuracy? Which of these metrics do you think the affect detection community should prioritize in the near- (next 5 years) and long- (next 15 years) term?