

## A Review and Meta-Analysis of Multimodal Affect Detection Systems

SIDNEY K. D'MELLO, University of Notre Dame

JACQUELINE KORY, MIT Media Lab

Affect detection is an important pattern recognition problem that has inspired researchers from several areas. The field is in need of a systematic review due to the recent influx of Multimodal (MM) affect detection systems that differ in several respects and sometimes yield incompatible results. This article provides such a survey via a quantitative review and meta-analysis of 90 peer-reviewed MM systems. The review indicated that the state of the art mainly consists of person-dependent models (62.2% of systems) that fuse audio and visual (55.6%) information to detect acted (52.2%) expressions of basic emotions and simple dimensions of arousal and valence (64.5%) with feature- (38.9%) and decision-level (35.6%) fusion techniques. However, there were also person-independent systems that considered additional modalities to detect nonbasic emotions and complex dimensions using model-level fusion techniques. The meta-analysis revealed that MM systems were consistently (85% of systems) more accurate than their best unimodal counterparts, with an average improvement of 9.83% (median of 6.60%). However, improvements were three times lower when systems were trained on natural (4.59%) versus acted data (12.7%). Importantly, MM accuracy could be accurately predicted (cross-validated  $R^2$  of 0.803) from unimodal accuracies and two system-level factors. Theoretical and applied implications and recommendations are discussed.

Categories and Subject Descriptors: I.5.m [Pattern Recognition]: Miscellaneous

General Terms: Measurement, Performance

Additional Key Words and Phrases: Affective computing, human-centered computing, evaluation, methodology, survey

### ACM Reference Format:

Sidney K. D'Mello and Jacqueline Kory. 2015. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*. ACM Comput. Surv. 47, 3, Article 43 (January 2015), 36 pages.

DOI: <http://dx.doi.org/10.1145/2682899>

## 1. INTRODUCTION

Affect detection (or affect recognition or affect classification) is an emerging research area of considerable practical and theoretical interest to a number of fields including signal processing, machine learning, computational linguistics, computer vision, neuroscience, and cognitive and social psychology [Picard 2010]. From a practical

This work is supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958) and the Bill & Melinda Gates Foundation via grants to the first author and NSF Graduate Research Fellowship under 1122374 to the second author. Any opinions, findings and conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the funding agencies.

Authors' addresses: S. K. D'Mello, Computer Science and Psychology at the University of Notre Dame, Notre Dame, IN 46556; email: [sdmello@nd.edu](mailto:sdmello@nd.edu); J. Kory, MIT Media Lab, Cambridge, MA 02139; email: [jakory@media.mit.edu](mailto:jakory@media.mit.edu).

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or [permissions@acm.org](mailto:permissions@acm.org).

© 2015 ACM 0360-0300/2015/01-ART43 \$15.00

DOI: <http://dx.doi.org/10.1145/2682899>

standpoint, affect detection is a cornerstone of affect-aware interfaces that aim to automatically detect and intelligently respond to users' affective states in order to increase usability and effectiveness [Brave and Nass 2002; Picard 1997]. From a theoretical standpoint, affect detection is ultimately a signal processing and pattern recognition problem because it involves the development of a classifier or regressor to detect an ill-defined phenomenon (affect) from observable signals. The problem is extremely challenging because affective states are psychological constructs (conceptual variables) that are not directly observable and are embedded in a noisy context-sensitive expressive and communicative system that has been fine-tuned over millions of years. The challenge is to detect an elusive and fleeting signal (affect) embedded in a system with multiple sources of noise exacerbated by context sensitivity, social masking, and individual and cultural variability [Elfenbein and Ambady 2002; Russell 1994; Russell et al. 2003].

The aforementioned complexities make affect detection an interesting and worthwhile problem to pursue as witnessed by numerous efforts toward detecting affective states from a variety of modalities, such as facial expressions, acoustic-prosodic cues, body movements, gesture, contextual cues, text and discourse, physiology, and neural circuitry (see Calvo and D'Mello [2010], Pantic and Rothkrantz [2003], and Zeng et al. [2009] for reviews). While early affect detection systems focused primarily on individual modalities and on emotional expressions portrayed by actors, many contemporary systems emphasize Multimodal (MM) detection of naturalistic affective expressions [Zeng et al. 2009], which is a novel problem in its own right.

Despite the impressive progress made so far, it is safe to say that there is still considerable ground to be covered before affect detectors can be integrated into everyday interfaces and devices and can be more readily deployed into real-world contexts. The field is still confronted with a number of persistent problems, such as (a) intrusive, expensive, and noisy sensors, some of which have scalability concerns; (b) technical challenges associated with detecting latent psychological constructs (i.e., affect) from weak signals embedded in noisy channels; (c) difficulties associated with collecting adequate realistic training data for machine learning [Douglas-Cowie et al. 2007]; (d) the persistent problem of obtaining ground truth labels for supervised classification, when interobserver agreement is generally low [Afzal and Robinson 2011; Graesser et al. 2006]; (e) challenges of incorporating top-down models of context with bottom-up body-based sensing [Conati and Maclaren 2009]; (f) issues of generalizability across contexts, time, individuals, and cultures [Calvo and D'Mello 2010]; (g) lack of clarity of the affective phenomenon being modeled (e.g., moods vs. emotions, categorical vs. dimensional representations, partly due to a difficulty in defining affect [Izard 2010]); and (g) many others as articulated in previous reviews [Calvo and D'Mello 2010; Pantic and Rothkrantz 2003; Zeng et al. 2009].

As researchers are well aware, this daunting list of challenges and open problems is more the norm than the exception given the difficulty of affect detection and the relative infancy of the field (about 15 years old). Numerous innovative solutions to address some of the aforementioned challenges have been extensively reviewed in both early (prior to 2009—see Cowie et al. [2001], Jaimes and Sebe [2007], Pang and Lee [2008], and Pantic and Rothkrantz [2003]) and more recent surveys (2009 to present—see Calvo and D'Mello [2010], D'Mello and Kory [2012], Valstar et al. [2012], and Zeng et al. [2009]), and will not be repeated here. Instead, the present focus is on *MM affect detection*, a strategy that is gaining momentum because it is expected to yield several advantages over unimodal (UM) affect detection. The remainder of the section briefly introduces the area of MM affect detection along with an overview of the issues addressed in this article.

**1.1. MM Affect Detection**

83

While UM detection involves the use of a single modality (e.g., facial features, gestures), MM systems fuse two or more modalities for affect detection. This raises a number of unique challenges and opportunities. The main challenges include (a) deciding which modalities to combine; (b) collecting MM training data; (c) handling missing data, different sampling rates, and modality interdependence when building models; (d) deciding how to fuse data from different modalities; and (e) deciding how to evaluate MM affect detectors. The hypothesized advantages of MM approaches to affect detection include (a) a higher-fidelity model of human affective expression, (b) a potential solution to address missing data caused by UM sensors, and (c) a solution to the noisy channel problem that plagues UM approaches.

84

85

86

87

88

89

90

91

92

93

With respect to the first advantage, it is widely acknowledged that human affective expression consists of a complex coordination of signals encompassing mostly involuntary (e.g., physiology), semivoluntary (facial expressions, body movements), and voluntary (e.g., overt actions such as key presses) responses [Ekman 1992; Rosenberg and Ekman 1994]. Analyzing multiple signals and their mutual interdependence is expected to yield models that more accurately reflect the underlying nature of human affective expression.

94

95

96

97

98

99

100

Second, UM signals suffer from notable problems associated with missing data. For example, a speech-based affect detector is virtually useless when the user is not speaking, while facial expressions cannot be reliably tracked when the face is out of view or occluded. MM approaches can provide more continuous affect detection capabilities by basing their decisions on the available channels.

101

102

103

104

105

The third hypothesized advantage of MM systems stems from the fact that UM affect detectors are inherently noisy since the link between specific signals and affective states is tenuous at best [Barrett et al. 2007; Russell et al. 2003]. This is partially the case because there is no one-to-one mapping between an expression and an affective state. For example, a furrowed brow caused by squinting to focus at something in the distance is diagnostic of a different cognitive state (information seeking) than a furrowed brow that accompanies an expression of confusion [D’Mello and Graesser 2014]. Furthermore, the same affective state can be differentially expressed as a function of the underlying eliciting stimulus. For example, a nearby spider (about to strike) and a spider across the room elicit different responses because they require different actions even though the underlying affective state (fear) elicited by both situations might be the same [Coan 2010]. In general, there is a loose coupling between observable expressions and specific affective states; hence, UM affect detectors are expected to yield moderate accuracies at best. MM affect detectors should yield improvements over UM systems because they are more suited to modeling the weak coupling between expression and experience of affect.

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

**1.2. Goals and Overview of the Present Article**

122

It is generally expected that incorporating MM signals should yield improvements in affect detection accuracies over UM signals. Although this assumption has obvious face validity, it has not always been supported. For example, when compared to the accuracies obtained by the best UM classifiers, some studies have reported impressive MM improvements (e.g., Jiang et al. [2011], Kessous et al. [2010], Lin et al. [2012], Paleari et al. [2009], and Wöllmer et al. [2010]), others have reported negligible or null improvements (e.g., Emerich et al. [2009], Kim [2007], and Metallinou et al. [2012]), and some have even reported negative effects (e.g., Glodek et al. [2011], Gunes and Piccardi [2005], and Khalali and Moradi [2009]). The considerable interstudy variance

123

124

125

126

127

128

129

130

131

132 in the results of MM affect detection makes it difficult to appropriately gauge what  
133 advantages (if any) MM detection yields over UM detection. In addition, there is the  
134 question of whether situations can be identified where MM detectors yield impressive  
135 improvements, and whether these situations can be differentiated from those that re-  
136 sult in null or negative effects. The present article attempts to address these questions  
137 by analyzing 90 MM and UM affect detection accuracies reported in published studies.

138 *Research Questions.* We focus on answering three specific research questions pertain-  
139 ing to state-of-the-art MM affect detection systems. First, what are the major trends  
140 in contemporary MM affect detectors? More specifically, can any general conclusions  
141 be drawn with respect to the various components (called *system-level factors*) of MM  
142 affect detection systems (e.g., type of training data, modality fusion methods, affect  
143 representation models)? Second, what is the added improvement (if any) in MM over  
144 the best UM detection accuracy (called MM1 effect size or MM1 effects)? Third, can we  
145 identify system-level factors that correlate with MM1 effects and can they be used to  
146 predict MM accuracies in a manner that generalizes across our sample of 90 studies  
147 (called moderation analyses)?

148 *Preliminary Analyses.* We have made an initial attempt to answer some of these ques-  
149 tions (specifically the second and partially the first and third questions) by performing  
150 a preliminary analysis of 30 published MM affect detectors [D'Mello and Kory 2012].  
151 The results of this initial analysis indicated that MM accuracies were consistently (26  
152 out of 30 studies) better than UM accuracies, and on average, yielded an 8.12% im-  
153 provement over the best UM detectors. The present article substantially expands on  
154 this initial study, both in terms of distributive breadth (the number of studies analyzed)  
155 and analysis depth (the types of questions that can be answered with a larger sample  
156 of studies).

157 *Focus of Current Analyses.* The focus of this article is on quantifying study-level  
158 factors and statistically analyzing MM accuracies rather than qualitatively describing  
159 individual affect detection systems; the latter has been extensively done in previous  
160 surveys, although mainly on UM and/or audio-visual detection (see Calvo and D'Mello  
161 [2010], Jaimes and Sebe [2007], Pantic and Rothkrantz [2003], and Zeng et al. [2009]).  
162 Hence, we do not discuss individual systems and approaches in depth, but focus on  
163 identifying general trends across systems with descriptive statistics and analyzing  
164 MM accuracies and effects with both descriptive and inferential statistics.

165 It is sometimes argued that meta-analyses of this type are not feasible because it is  
166 improper to compare accuracies across studies that differ in multiple respects. Hence,  
167 it is important to emphasize that the present article does not make such comparisons.  
168 Instead, MM1 effects are computed by comparing MM accuracies to UM accuracies  
169 from the *same* study, a comparison that is justifiable because study-level factors are  
170 held constant. The distribution of MM1 effects from individual studies is then statisti-  
171 cally analyzed, an approach recommended by standard texts on meta-analyses (e.g.,  
172 Borenstein et al. [2009] and Lipsey and Wilson [2001]). In addition, the variability in  
173 datasets, methods, and metrics used is, in fact, a major strength of meta-analytical ap-  
174 proaches because it allows one to estimate “population effects” from individual “study  
175 effects” by averaging across interstudy variability.

176 To summarize, with the exception of our preliminary study [D'Mello and Kory 2012],  
177 this article represents the first major attempt to quantify and statistically analyze a  
178 large set of MM affect detectors in order to make generalizable conclusions.

## 179 2. METHOD

180 The methodology used to search for relevant articles, the inclusion/exclusion criteria,  
181 the data coding, and data treatment procedures are discussed in some detail in this  
182 section to enable replication as more studies emerge in the literature.

## 2.1. Search Process and Inclusion/Exclusion Criteria

183

A three-pronged approach was used for study selection. First, relevant journals and conference proceedings were searched using a *targeted search* strategy. The journals included *IEEE Transactions on Affective Computing*, *IEEE Transactions on Multimedia*, and *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Conferences included the *International Conference on Affective Computing and Intelligent Interaction (ACII)*, *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, *IEEE International Conference on Multimedia and Expo (ICME)*, *ACM International Conference on Multimodal Interfaces (ICMI)*, and *INTER\_SPEECH*. The *secondary search* commenced by identifying additional articles from the reference sections of articles retrieved from the targeted search and from recent survey articles [Calvo and D’Mello 2010; Zeng et al. 2009]. Finally, the *informal search* proceeded by querying Google Scholar with the following search queries: (multimodal OR bimodal) fusion; (affect OR emotion) AND (detection OR recognition). We restricted our targeted search to articles published within the last 5 years (2009–2013), but earlier articles could have been retrieved in the secondary and informal searches as long as they were published in the last 10 years (2003 and beyond).

184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199

A rather liberal inclusion/exclusion criterion was adopted in order to maximize the number of studies considered. Any peer-reviewed publication that reported both UM and MM affect detection accuracies in a clearly accessible format (i.e., accuracy metrics could be easily obtained from the text, tables, or figures) was included in the analysis. Failure to report both UM and MM accuracies unfortunately led to the exclusion of some relevant and highly cited studies (e.g., Kapoor et al. [2007]), but this was unavoidable due to the nature of the analytic strategy. Selection bias was avoided by never excluding a study based on the results, publication outlet, or authors.

200  
201  
202  
203  
204  
205  
206  
207

In all, 84 articles were selected based on the search and inclusion/exclusion criteria. These 84 articles yielded 90 viable systems since some articles reported more than one unique multimodal affect detector. There was a strong positive correlation between the year (2004–2013) and the number of studies,  $r = 0.727$ , suggesting that recent studies were more frequent in the sample. More than 60% of the studies were from the 2009–2013 period and 42% of all studies were from the 2011–2013 period.

208  
209  
210  
211  
212  
213

## 2.2. Data Coding

214

The studies were coded along several system-level (or study-level) factors. The coding process was initially performed by one of the authors and then independently checked by the second author. Disagreements were resolved via discussion among the authors. Table I describes how each study was coded with respect to the factors discussed in the following.

215  
216  
217  
218  
219

*Data type* addresses whether training and validation data consisted of affective expressions that were (a) obtained by asking *actors* to portray various emotions (e.g., Castellano et al. [2008], Cueva et al. [2011], Dobrišek et al. [2013], Lingenfelser et al. [2011], and Metallinou et al. [2012]), (b) collected via experimental methods that *induced* specific emotions (e.g., Bailenson et al. [2008], Glodek et al. [2013], Koelstra et al. [2012], Soleymani et al. [2012], and Wöllmer et al. [2013a]), or (c) *naturalistic* displays of affect (i.e., nonacted and not induced—e.g., Castellano et al. [2009], D’Mello and Graesser [2010], Kapoor and Picard [2005], Litman and Forbes-Riley [2004], and Wöllmer et al. [2013b]).

220  
221  
222  
223  
224  
225  
226  
227  
228

While the criteria for a dataset to be categorized as acted or natural is quite clear, the induced category requires some clarification. This designation was applied to datasets where specific emotions were induced using well-established techniques such as showing participants films (e.g., Soleymani et al. [2012]) or images (e.g., Hussain et al.

229  
230  
231  
232

Table 1. Details of Individual Study Characteristics (Sorted by Author and Year)

| Reference                      | N  | Data Type | Rep. Model | Class Model | k | Affect States | Modalities            | Fusion Type | Validation Method |
|--------------------------------|----|-----------|------------|-------------|---|---------------|-----------------------|-------------|-------------------|
| Ballenson et al. [2008]        | 41 | ind       | disc       | reg         |   | disc mixed    | Face + PPhy           | feat        | indep             |
| Baltrušaitis et al. [2013]     | 16 | ind       | dim        | reg         |   | dim complex   | Face + Voice          | model       | indep             |
| Banda and Robinson [2011]      | 4  | act       | disc       | class       | 7 | disc basic    | Face + Voice          | dec         | dep               |
| Busso et al. [2004]            | 1  | act       | disc       | class       | 4 | disc basic    | Face + Voice          | feat        | dep               |
| Caridakis et al. [2006]        | 4  | ind       | dim        | class       | 4 | dim simple    | Face + Voice          | model       | dep               |
| Castellano et al. [2008]       | 10 | act       | disc       | class       | 8 | disc mixed    | Face + Voice + Body   | feat        | dep               |
| Castellano et al. [2009]       | 8  | nat       | disc       | class       | 2 | disc nonbasic | Face + Content        | feat        | indep             |
| Chanel et al. [2011]           | 20 | nat       | disc       | class       | 3 | disc mixed    | CPhy + PPhy           | dec         | indep             |
| Chen et al. [2005]             | 2  | act       | disc       | class       | 7 | disc basic    | Face + Voice          | feat        | dep               |
| Chetty and Wagner [2008]       | 8  | act       | disc       | class       | 4 | disc basic    | Face + Voice          | hybrid      | dep               |
| Chetty and Wagner [2008]       | 44 | act       | disc       | class       | 4 | disc basic    | Face + Voice          | hybrid      | dep               |
| Chuang and Wu [2004]           | 2  | act       | disc       | class       | 7 | disc basic    | Voice + Text          | dec         | dep               |
| Cueva et al. [2011]            | 42 | act       | disc       | class       | 4 | disc basic    | Face + Voice          | dec         | dep               |
| Datu and Rothkrantz [2011]     | 42 | act       | disc       | class       | 6 | disc basic    | Face + Voice          | feat        | dep               |
| D'Mello and Graesser [2007]    | 28 | nat       | disc       | class       | 4 | disc nonbasic | Body + Content        | feat        | dep               |
| D'Mello and Graesser [2010]    | 28 | nat       | disc       | class       | 5 | disc nonbasic | Face + Body + Content | feat        | indep             |
| Dobrišek et al. [2013]         | 43 | act       | disc       | class       | 6 | disc basic    | Face + Voice          | dec         | dep               |
| Dy et al. [2010]               |    | nat       | disc       | class       | 5 | disc basic    | Face + Voice          | dec         | dep               |
| Emerich et al. [2009]          | 28 | act       | disc       | class       | 6 | disc basic    | Face + Voice          | feat        | dep               |
| Eyben et al. [2010]            | 4  | ind       | dim        | reg         |   | dim simple    | Voice + Text          | model       | dep               |
| Forbes-Riley and Litman [2004] | 17 | nat       | disc       | class       | 3 | dim simple    | Voice + Text          | feat        | dep               |
| Gajsek et al. [2010]           | 42 | act       | disc       | class       | 6 | disc basic    | Face + Voice          | dec         | dep               |
| Glodek et al. [2011]           | 16 | ind       | dim        | class       | 2 | dim complex   | Face + Voice          | dec         | indep             |
| Glodek et al. [2013]           | 16 | ind       | dim        | class       | 2 | dim complex   | Face + Voice          | dec         | indep             |
| Gong et al. [2007]             | 23 | act       | disc       | class       | 7 | disc mixed    | Face + Body           | feat        | dep               |
| Gunes and Piccardi [2005]      | 4  | act       | disc       | class       | 6 | disc mixed    | Face + Body           | feat        | dep               |

Continued

Table 1. Continued

| Reference                       | $N$ | Data Type | Rep. Model | Class Model | $k$ | Affect States | Modalities                    | Fusion Type | Validation Method |
|---------------------------------|-----|-----------|------------|-------------|-----|---------------|-------------------------------|-------------|-------------------|
| Gunes and Piccardi [2009]       | 10  | act       | disc       | class       | 12  | disc mixed    | Face + Body                   | feat        | dep               |
| Han et al. [2007]               | 14  | act       | disc       | class       | 5   | disc basic    | Face + Voice                  | dec         | dep               |
| Haq et al. [2008]               | 1   | act       | disc       | class       | 7   | disc basic    | Face + Voice                  | feat        | dep               |
| Haq and Jackson [2009]          | 4   | act       | disc       | class       | 7   | disc basic    | Face + Voice                  | dec         | dep               |
| Hoch et al. [2005]              | 7   | act       | disc       | class       | 4   | disc mixed    | Face + Voice                  | dec         | dep               |
| Hommel et al. [2013]            | 4   | act       | disc       | class       | 5   | disc basic    | Face + Voice                  | dec         | dep               |
| Hussain et al. [2012]           | 19  | ind       | dim        | class       | 3   | dim simple    | Face + PPhy                   | dec         | dep               |
| Jiang et al. [2011]             | 42  | act       | disc       | class       | 6   | disc basic    | Face + Voice                  | model       | dep               |
| Joo et al. [2007]               | 5   | act       | disc       | class       | 5   | disc basic    | Face + Voice                  | dec         | dep               |
| Kanluan et al. [2008]           | 20  | nat       | dim        | reg         |     | dim complex   | Face + Voice                  | dec         | dep               |
| Kapoor and Picard [2005]        | 8   | nat       | disc       | class       | 2   | disc nonbasic | Face + Body + Content         | model       | dep               |
| Karpouzism et al. [2007]        | 4   | ind       | dim        | class       | 4   | dim simple    | Face + Voice                  | feat        | dep               |
| Kessous et al. [2010]           | 10  | act       | disc       | class       | 8   | disc mixed    | Face + Voice                  | feat        | dep               |
| Khalali and Moradi [2009]       | 5   | ind       | disc       | class       | 3   | disc mixed    | Face + Body + PPhy            | feat        | dep               |
| Kim et al. [2005]               | 3   | ind       | dim        | class       | 4   | dim simple    | Voice + PPhy                  | feat        | dep               |
| Kim [2007]                      | 3   | ind       | dim        | class       | 4   | dim simple    | Voice + PPhy                  | feat        | indep             |
| Kim and Lingenfelter [2010]     | 3   | ind       | dim        | class       | 4   | dim simple    | Voice + PPhy                  | dec         | dep               |
| Koelstra et al. [2012]          | 22  | ind       | dim        | class       | 2   | dim complex   | Voice + PPhy + Content + PPhy | dec         | dep               |
| Krell et al. [2013]             | 13  | ind       | disc       | class       | 2   | disc nonbasic | Face + Voice                  | dec         | indep             |
| Lin et al. [2012]               | 7   | act       | disc       | class       | 4   | disc basic    | Face + Voice                  | model       | indep             |
| Lin et al. [2012]               | 4   | ind       | dim        | class       | 4   | dim simple    | Face + Voice                  | model       | indep             |
| Lingenfelter et al. [2011]      | 8   | act       | dim        | class       | 7   | disc basic    | Face + Voice                  | dec         | indep             |
| Lingenfelter et al. [2011]      | 13  | ind       | dim        | class       | 3   | dim simple    | Face + Voice                  | dec         | indep             |
| Litman and Forbes-Riley [2004]  | 15  | nat       | dim        | class       | 3   | dim simple    | Voice + Text                  | feat        | dep               |
| Litman and Forbes-Riley [2006a] | 14  | nat       | dim        | class       | 3   | dim simple    | Voice + Text                  | feat        | dep               |

Continued

Table I. Continued

| Reference                        | N  | Data Type | Rep. Model | Class Model | k  | Affect States | Modalities          | Fusion Type | Validation Method |
|----------------------------------|----|-----------|------------|-------------|----|---------------|---------------------|-------------|-------------------|
| Litman and Forbes-Riley [2006a]  | 20 | nat       | dim        | class       | 3  | dim simple    | Voice + Text        | feat        | dep               |
| Lu and Jia [2012]                | 5  | act       | dim        | class       | 2  | dim simple    | Face + Voice        | model       | indep             |
| Mansoorzadeh and Charkari [2010] | 12 | act       | disc       | class       | 6  | disc basic    | Face + Voice        | hybrid      | dep               |
| Mansoorzadeh and Charkari [2010] | 42 | act       | disc       | class       | 6  | disc basic    | Face + Voice        | hybrid      | dep               |
| Metallinou et al. [2008]         | 10 | act       | disc       | class       | 4  | disc basic    | Face + Voice        | dec         | dep               |
| Metallinou et al. [2012]         | 10 | act       | dim        | class       | 3  | dim simple    | Face + Voice        | model       | indep             |
| Monkarese et al. [2012]          | 20 | ind       | dim        | class       | 2  | dim simple    | Face + PPhy         | feat        | dep               |
| Nicolaou et al. [2011]           | 4  | ind       | dim        | reg         |    | dim simple    | Face + Voice + Body | model       | dep               |
| Pal et al. [2006]                |    | nat       | disc       | class       | 5  | disc mixed    | Face + Voice        | dec         | dep               |
| Paleari et al. [2009]            | 44 | act       | disc       | class       | 6  | disc basic    | Face + Voice        | model       | indep             |
| Park et al. [2012]               | 10 | act       | disc       | class       | 4  | disc basic    | Face + Voice        | dec         | dep               |
| Rabie et al. [2009]              | 8  | act       | disc       | class       | 7  | disc basic    | Face + Voice        | model       | indep             |
| Rashid et al. [2012]             | 42 | act       | disc       | class       | 6  | disc basic    | Face + Voice        | dec         | dep               |
| Rigoll et al. [2005]             | 13 | act       | disc       | class       | 7  | disc basic    | Voice + Text        | dec         | indep             |
| Rosas et al. [2013]              | 76 | nat       | dim        | class       | 2  | dim simple    | Face + Voice + Text | feat        | indep             |
| Rosas et al. [2013]              | 37 | nat       | dim        | class       | 2  | dim simple    | Face + Voice + Text | feat        | indep             |
| Rozgic et al. [2012]             | 10 | act       | disc       | class       | 4  | disc basic    | Face + Voice + Text | feat        | indep             |
| Savran et al. [2012]             | 16 | ind       | dim        | reg         |    | dim complex   | Face + Voice + Text | model       | indep             |
| Schuller et al. [2007]           | 21 | nat       | disc       | class       | 3  | disc nonbasic | Face + Voice        | feat        | indep             |
| Schuller [2011]                  | 47 | nat       | dim        | reg         |    | dim complex   | Voice + Text        | feat        | indep             |
| Sebe et al. [2006]               | 38 | act       | disc       | class       | 11 | disc mixed    | Face + Voice        | model       | dep               |
| Seppi et al. [2008]              | 51 | ind       | disc       | class       | 4  | disc mixed    | Voice + Text        | feat        | indep             |
| Shan et al. [2007]               | 5  | act       | disc       | class       | 7  | disc mixed    | Face + Body         | feat        | dep               |
| Soleymani et al. [2012]          | 24 | ind       | dim        | class       | 3  | dim simple    | CPhy + Gaze         | dec         | indep             |
| Tu and Yu [2012]                 | 42 | act       | disc       | class       | 6  | disc basic    | Face + Voice        | dec         | dep               |

Continued



Table I. Continued

| Reference              | $N$ | Data Type | Rep. Model | Class Model | $k$ | Affect States | Modalities          | Fusion Type | Validation Method |
|------------------------|-----|-----------|------------|-------------|-----|---------------|---------------------|-------------|-------------------|
| Vu et al. [2011]       | 5   | act       | disc       | class       | 4   | disc mixed    | Voice + Body        | dec         | dep               |
| Wagner et al. [2011]   | 21  | act       | dim        | class       | 4   | dim simple    | Face + Voice + Body | dec         | indep             |
| Walter et al. [2011]   | 10  | ind       | dim        | class       | 2   | dim complex   | Voice + PPhy        | dec         | dep               |
| Wang and Guan [2005]   | 8   | act       | disc       | class       | 6   | disc basic    | Face + Voice        | feat        | indep             |
| Wang and Guan [2008]   | 8   | act       | disc       | class       | 6   | disc basic    | Face + Voice        | feat        | indep             |
| Wang et al. [2013]     | 28  | ind       | dim        | class       | 2   | dim simple    | CPhy + Content      | feat        | dep               |
| Wimmer et al. [2008]   | 8   | ind       | disc       | class       | 6   | disc nonbasic | Face + Voice        | feat        | dep               |
| Wöllmer et al. [2010]  | 10  | act       | dim        | class       | 3   | dim simple    | Face + Voice        | feat        | indep             |
| Wöllmer et al. [2013a] | 16  | ind       | dim        | class       | 2   | dim complex   | Face + Voice        | model       | indep             |
| Wöllmer et al. [2013b] | 343 | nat       | dim        | class       | 2   | dim simple    | Face + Voice + Text | hybrid      | indep             |
| Wu and Liang [2011]    | 8   | act       | disc       | class       | 4   | disc basic    | Voice + Text        | dec         | dep               |
| Zeng et al. [2005]     | 20  | act       | disc       | class       | 11  | disc mixed    | Face + Voice        | model       | indep             |
| Zeng et al. [2006]     | 2   | nat       | disc       | class       | 2   | dim simple    | Face + Voice        | model       | dep               |
| Zeng et al. [2007]     | 20  | act       | disc       | class       | 11  | disc mixed    | Face + Voice        | model       | indep             |

Notes:  $N$  = number of participants (blank when not specified); Data Type (act = acted; ind = induced; nat = natural); Representation Model (disc = discrete; dim = dimensional); Classification Model (class = classification, reg = regression);  $k$  = Number of affective states (only for classification tasks, otherwise blank); Affect states (disc basic = discrete basic emotions; disc nonbasic = discrete nonbasic emotions; disc mixed = discrete basic + nonbasic emotions; dim simple = dimensional simple; dim complex = dimensional complex); Modalities (PPhy = peripheral physiology, CPhy = central physiology; Content = content/context); Fusion Type (feat = feature; dec = decision); Validation Method (dep = subject dependent; indep = subject independent).

233 [2012]) that were previously validated as being reliable elicitors of affect [Kory and  
234 D'Mello 2014]. It was also applied to studies where individuals were required to par-  
235 ticipate in interactions that were intentionally affectively charged, thereby increas-  
236 ing the likelihood that they would respond emotionally. For example, the SEMAINE  
237 dataset [McKeown et al. 2012] was constructed by asking individuals to engage in a  
238 conversation with an animated agent that had one of four affective dispositions (or  
239 personalities): angry, happy, gloomy, or pragmatic. Studies that utilized this dataset  
240 (e.g., Karpouzis et al. [2007] and Nicolaou et al. [2011]) were categorized as “induced”  
241 because it is likely that the affective disposition of the agent induced specific emotions  
242 in the individual. In fact, this was the main motivation toward using agents with four  
243 specific affective dispositions.

244 *Number of participants* simply refers to the number of unique individuals in the  
245 training/validation dataset. It is an important factor because generalizability is related  
246 to the number of individuals used to train the detector due to individual differences in  
247 affect expression.

248 *Affect representation model* refers to whether ground truth affect measures for the  
249 supervised classifiers consisted of *discrete* or *dimensional* representations. Discrete  
250 models consider emotional episodes as belonging to one of  $m$  distinct categories (e.g.,  
251 judging if a 30 second video of an individual's face represents anger, sadness, or fear).  
252 Discrete ratings do not need to be mutually exclusive since affective blends are often  
253 experienced, yet most studies use mutually exclusive ratings for convenience (e.g.,  
254 D'Mello and Graesser [2010], Krell et al. [2013], and Rashid et al. [2012]). Dimensional  
255 models represent affect along one or more dimensions, primarily valence (positive-  
256 negative) and activation/arousal (sleepy vs. awake or inactive vs. active) (e.g., Hussain  
257 et al. [2012], Lu and Jia [2012], and Wang et al. [2013]), but occasionally extending to  
258 other dimensions such as expectancy, power, and dominance (e.g., Baltrušaitis et al.  
259 [2013], Glodek et al. [2013], and Wöllmer et al. [2013a]).

260 The affect representation model is a conceptual entity that is concerned with the  
261 affective representation and not with the measurement scale per se. Hence, studies  
262 involving ordinal or continuous ratings of discrete emotions were coded as discrete,  
263 as was the case where the intensity of amusement (a discrete state) was rated via a  
264 0 (neutral) to 8 (amused) scale (e.g., Bailenson et al. [2008]). Similarly, studies with  
265 categorical ratings of dimensions (e.g., low vs. high ratings of valence) were coded as  
266 dimensional (e.g., Bailenson et al. [2008]).

267 *Affect detection model* pertains to whether the machine learning models were clas-  
268 sifiers or regressors. In most cases, classifiers and regressors were used when affect  
269 models were discrete (e.g., D'Mello and Graesser [2010], Hommel et al. [2013], and  
270 Rashid et al. [2012]) and continuous (e.g., Eyben et al. [2011], Kanluan et al. [2008],  
271 and Savran et al. [2012]), respectively. However, a number of studies used dimensional  
272 representations and collected ordinal or continuous ratings, but performed classifica-  
273 tions instead of regressions by discretizing the scales into high versus low or high versus  
274 medium versus low categories (e.g., Glodek et al. [2011] and Wöllmer et al. [2013a]).  
275 For example, Wöllmer et al. [2010] used a five-point scale to measure valence and  
276 activation, but then performed a categorical classification by performing a tripartite  
277 split on each dimension (i.e., dividing the scale into low, medium, and high sections).  
278 Similarly, ordinal or continuous activation-valence values were often discretized by  
279 clustering prior to classification (e.g., Karpouzis et al. [2007]).

280 *Number of affective states detected* only applies to classification tasks and is sim-  
281 ply the number of discrete affective states considered. It is an important factor as  
282 the affect detection problem ostensibly becomes more challenging as the number of  
283 discriminations increases.

*Affective states/dimensions detected* pertains to the specific affective states/dimensions in the classification/regression models. Researchers in the affective sciences have proposed a number of taxonomies to categorize the discrete affective states that occur in everyday experiences [Ekman 1992; Ortony et al. 1988; Plutchik 2001]. Broadly, the affective states can be divided into *discrete basic* and *discrete nonbasic* states. States such as anger, surprise, happiness, disgust, sadness, and fear are typically considered to be basic affective states [Ekman 1992]. States such as boredom, confusion, frustration, engagement, and curiosity share some, but not all, of the features commonly attributed to basic emotions (see Ekman [1992]). Consequently, these are labeled as *nonbasic* states. Some studies used a combination of both (e.g., Castellano et al. 2008; Sebe et al. 2006) and these were coded as *discrete mixed*.

With respect to affective dimensions, most researchers agree that valence and arousal (activation) are two essential dimensions to represent affect [Barrett et al. 2007; Russell 2003]. Beyond this, there is considerable debate as to which other dimensions are needed [Fontaine et al. 2007; Kaernbach 2011]. Most studies detected valence and arousal (coded as *dimensional simple*), but expectancy, power, and dominance were also considered in some studies (coded as *dimensional complex*).

*Number of modalities* simply refers to whether the MM detectors fused two (*bimodal*) or three (*trimodal*) modalities.

*Modalities* refer to the specific modalities used for affect detection. In communication theory, *modality* is considered to be distinct from *medium* because the former focuses on the sense via which a message is communicated (e.g., facial expression, pitch), while the latter is concerned with the means of message communication [Sutcliffe 2008]. For example, facial expressions and gestures are different modalities that can be communicated via the same medium (video). The present coding scheme focused on modality instead of medium.

The specific modalities used in the 90 studies included (a) *facial* features extracted from video, (b) paralinguistic or acoustic-prosodic features from the *voice*, (c) linguistic or semantic features from written or spoken *language*, (d) *body movements* consisting of postures and gestures (excluding facial features), (e) *eye gaze*, (f) *central physiology* (only Electroencephalography—EEG), (g) *peripheral physiology* (e.g., Electrodermal activity (EDR), Electrocardiography (ECG), Electromyography (EMG), respiration), and (h) *content and context*.

While modalities (a)–(f) were straightforward, peripheral physiology and content/context require some clarification. With respect to peripheral physiology, although individual channels, such as EDR, ECG, EMG, and so forth, can be analyzed independently and treated as separate modalities, most studies fused features from these various channels instead of considering each signal individually. For example, Chanel et al. [2011] built (a) a peripheral model by combining galvanic skin response, blood volume pulse, heart rate, chest cavity expansion, and skin temperature; (b) a central physiology model (EEG); and (c) a combined peripheral + central physiology model. In this and similar cases, the combination of the individual peripheral physiological channels was taken as a UM detector.

Content features were gleaned from a multimedia content analysis of affect-elicitation stimuli (e.g., low-level video features such as color, lighting [Koelstra et al. 2012]). Context features were obtained by analyzing the situation in which the affective interaction was embedded. For example, D’Mello and Graesser [2010] tracked a number of contextual cues, such as session length, system feedback, and so on, when individuals completed a learning session with a computer tutor. Both content and context features are unique from the other modalities in that they are obtained from the stimuli and situation rather than the individuals themselves. They were grouped as

335 context/context features since there were not a sufficient number of studies to sustain  
 336 an independent analysis of each.

337 *Fusion method* pertains to the method used to fuse modalities. Possible options  
 338 include *data-level*, *decision-level*, *score-level*, *hybrid*, and *model-level* fusion. In data-  
 339 level fusion, individual data streams are fused prior to feature engineering (e.g., fusing  
 340 video data from two cameras). Feature-level fusion consists of independently computing  
 341 features from each modality and then fusing the features prior to classification (e.g.,  
 342 Castellano et al. [2008], D'Mello and Graesser [2010], and Litman and Forbes-Riley  
 343 [2006a]). In decision-level fusion, classification is first performed on the individual  
 344 features and the outputs (decisions) are fused via one of several voting rules (e.g.,  
 345 Kanluan et al. [2008], Koelstra et al. [2012], and Walter et al. [2011]). Score-level  
 346 fusion is related to decision-level fusion in that affect likelihoods (or probabilities)  
 347 computed by classifiers operating on independent modalities are fused (e.g., Gajsek  
 348 et al. [2010]). Only a small number of systems relied on score-level fusion, so these were  
 349 coded as decision-level fusion due to the similarity between these two methods. Hybrid  
 350 fusion combines both feature- and decision-level fusion, for example, by combining  
 351 independent decisions of individual UM classifiers with the decisions of a feature-  
 352 level fused MM classifier (e.g., Chetty and Wagner [2008] and Mansoorizadeh and  
 353 Charkari [2010]). Finally, model-level fusion takes advantage of the interdependencies  
 354 among the various modalities during the fusion process (e.g., Caridakis et al. [2006],  
 355 Eyben et al. [2010], and Metallinou et al. [2012]). When multiple fusion techniques  
 356 were implemented and compared in a single study, the fusion method that yielded the  
 357 highest accuracy was analyzed.

358 *Validation method* is concerned with whether the affect detectors are expected to  
 359 generalize to new individuals (person independent) or not (person dependent). This  
 360 is a critical distinction because (for the most part) affect detectors are intended to  
 361 be person independent but developing such systems is more challenging due to large  
 362 interindividual variability in affect. Designation of an affect detector as person depen-  
 363 dent or independent was rarely articulated in the papers, but could be inferred from  
 364 the methods used to validate the detectors. Studies that used leave-one-person-out or  
 365 leave-several-people-out validation techniques, where instances from the same individ-  
 366 ual were either in the training or testing sets but never both, were deemed to be *person*  
 367 *independent* (e.g., D'Mello and Graesser [2010], Savran et al. [2012], and Schuller  
 368 [2011]). Studies that cross-validated within an individual, or studies where person in-  
 369 dependence across training and testing sets was not carefully controlled were coded as  
 370 *person dependent* (e.g., Castellano et al. [2008], Litman and Forbes-Riley [2006a], and  
 371 Monkaresi et al. [2012]).

### 372 2.3. Encoding Affect Detection Accuracy

373 Table II provides several measures of UM and MM affect detection accuracies. The  
 374 key measures were detection accuracy of the best, second-best, and worst UM detec-  
 375 tors (Max1, Max2, and Min, respectively) and MM accuracy (MM). Most studies that  
 376 performed a categorical classification used classification accuracy (i.e., the proportion  
 377 of correctly classified instances) as the evaluation metric. In rare cases where both  
 378 classification accuracy and the F1 measure were reported, classification accuracy was  
 379 taken to be the metric in order to increase consistency among studies. The correlation  
 380 coefficient was taken as the performance metric for regression models.

381 MM1 effect was the key effect size metric. If  $a_1$  and  $a_2$  are accuracies associated with  
 382 two UM detectors, and  $a_{12}$  is the MM accuracy, then the MM1 effect was computed as  
 383 the percent improvement over the best UM detector (see Equation (1)). This metric  
 384 affords a unified analysis framework for studies that used classification accuracies, F1

Table II. Classification Accuracy and Multimodal (MM) Effect Sizes (Sorted by Author + Year)

| Reference                      | Meas. | UM    |       |       | Text  | Body | Gaze  | PPhy  | CPhy  | Content | MM    | MM Effect Size (%) |       |       |
|--------------------------------|-------|-------|-------|-------|-------|------|-------|-------|-------|---------|-------|--------------------|-------|-------|
|                                |       | Face  | Voice | MM1   |       |      |       |       |       |         |       | MM2                | MMMin |       |
| Bailenson et al. [2008]        | CC    | 0.405 |       |       |       |      | 0.280 |       |       |         | 0.440 | 15.2               | 147.9 | 147.9 |
| Baltrušaitis et al. [2013]     | CC    | 0.248 | 0.201 |       |       |      |       |       |       |         | 0.301 | 1.6                | 154.8 | 154.8 |
| Banda and Robinson [2011]      | Acc   | 0.952 | 0.791 |       |       |      |       |       |       |         | 0.977 | 2.6                | 23.5  | 23.5  |
| Busso et al. [2004]            | Acc   | 0.851 | 0.709 |       |       |      |       |       |       |         | 0.891 | 4.7                | 25.7  | 25.7  |
| Caridakis et al. [2006]        | Acc   | 0.670 | 0.730 |       |       |      |       |       |       |         | 0.790 | 8.2                | 17.9  | 17.9  |
| Castellano et al. [2008]       | Acc   | 0.483 | 0.571 |       | 0.671 |      |       |       |       |         | 0.783 | 16.7               | 37.1  | 62.1  |
| Castellano et al. [2009]       | Acc   | 0.938 |       |       |       |      |       |       | 0.781 |         | 0.948 | 1.1                | 21.4  | 21.4  |
| Chen et al. [2011]             | Acc   |       |       |       |       |      | 0.590 | 0.560 |       |         | 0.630 | 6.8                | 12.5  | 12.5  |
| Chen et al. [2005]             | Acc   | 0.750 | 0.630 |       |       |      |       |       |       |         | 0.840 | 12.0               | 33.3  | 33.3  |
| Chetty and Wagner [2008]       | Acc   | 0.850 | 0.709 |       |       |      |       |       |       |         | 0.970 | 14.1               | 36.8  | 36.8  |
| Chetty and Wagner [2008]       | Acc   | 0.820 | 0.644 |       |       |      |       |       |       |         | 0.964 | 17.6               | 49.7  | 49.7  |
| Chuang and Wu [2004]           | Acc   |       | 0.764 | 0.655 |       |      |       |       |       |         | 0.815 | 6.7                | 24.4  | 24.4  |
| Cueva et al. [2011]            | Acc   | 0.200 | 0.650 |       |       |      |       |       |       |         | 0.750 | 15.4               | 275.0 | 275.0 |
| Datu and Rothkrantz [2011]     | Acc   | 0.377 | 0.559 |       |       |      |       |       |       |         | 0.563 | 0.7                | 49.3  | 49.3  |
| D'Mello and Graesser [2007]    | Acc   | 0.352 |       |       | 0.331 |      |       |       |       | 0.391   | 0.407 | 4.1                | 23.0  | 23.0  |
| D'Mello and Graesser [2010]    | Acc   | 0.528 | 0.725 |       | 0.316 |      |       |       |       | 0.381   | 0.487 | 6.8                | 43.7  | 98.3  |
| Dobrišek et al. [2013]         | Acc   | 0.860 | 0.400 |       |       |      |       |       |       |         | 0.775 | 6.9                | 46.8  | 46.8  |
| Dy et al. [2010]               | Acc   | 0.907 | 0.877 |       |       |      |       |       |       |         | 0.800 | -7.0               | 100.0 | 100.0 |
| Emerich et al. [2009]          | CC    |       | 0.505 | 0.370 |       |      |       |       |       |         | 0.930 | 2.5                | 6.0   | 6.0   |
| Eyben et al. [2010]            | CC    |       | 0.762 | 0.832 |       |      |       |       |       |         | 0.530 | 5.2                | 45.2  | 45.2  |
| Forbes-Riley and Litman [2004] | Acc   |       | 0.629 |       |       |      |       |       |       |         | 0.837 | 0.6                | 9.8   | 9.8   |
| Gajsek et al. [2010]           | Acc   | 0.547 | 0.629 |       |       |      |       |       |       |         | 0.713 | 13.4               | 30.3  | 30.3  |
| Glodek et al. [2011]           | Acc   | 0.500 | 0.506 |       |       |      |       |       |       |         | 0.470 | -14.2              | 8.3   | 8.3   |
| Glodek et al. [2013]           | Acc   | 0.620 | 0.620 |       |       |      |       |       |       |         | 0.643 | 0.3                | 8.0   | 8.0   |
| Gong et al. [2007]             | Acc   | 0.809 |       |       | 0.746 |      |       |       |       |         | 0.896 | 10.8               | 20.1  | 20.1  |
| Gunes and Piccardi [2005]      | Acc   | 0.829 |       |       | 1.000 |      |       |       |       |         | 1.000 | 0.0                | 20.6  | 20.6  |
| Gunes and Piccardi [2009]      | Acc   | 0.352 |       |       | 0.769 |      |       |       |       |         | 0.827 | 7.5                | 134.9 | 134.9 |
| Han et al. [2007]              | Acc   | 0.817 | 0.737 |       |       |      |       |       |       |         | 0.869 | 6.4                | 17.9  | 17.9  |
| Haq et al. [2008]              | Acc   | 0.983 | 0.525 |       |       |      |       |       |       |         | 0.983 | 0.0                | 87.2  | 87.2  |
| Haq and Jackson [2009]         | Acc   | 0.954 | 0.563 |       |       |      |       |       |       |         | 0.975 | 2.2                | 73.2  | 73.2  |
| Hoch et al. [2005]             | Acc   | 0.668 | 0.868 |       |       |      |       |       |       |         | 0.907 | 4.5                | 35.8  | 35.8  |
| Hommel et al. [2013]           | Acc   | 0.554 | 0.236 |       |       |      |       |       |       |         | 0.581 | 5.0                | 146.0 | 146.0 |

Continued

Table II. Continued

| Reference                         | Meas. | UM    |       |    |       | Text | Body | Gaze  | PPhy  | CPhy  | Content | MM Effect Size (%) |      |       |       |
|-----------------------------------|-------|-------|-------|----|-------|------|------|-------|-------|-------|---------|--------------------|------|-------|-------|
|                                   |       | Face  | Voice | UM | MM    |      |      |       |       |       |         | MM1                | MM2  | MMMin |       |
| Hussain et al. [2012]             | Acc   | 0.585 |       |    |       |      |      | 0.495 |       |       |         | 0.624              | 6.5  | 25.9  | 25.9  |
| Jiang et al. [2011]               | Acc   | 0.468 | 0.522 |    |       |      |      |       |       |       |         | 0.665              | 27.4 | 42.1  | 42.1  |
| Joo et al. [2007]                 | Acc   | 0.534 | 0.630 |    |       |      |      |       |       |       |         | 0.704              | 11.7 | 31.8  | 31.8  |
| Kanluan et al. [2008]             | CC    | 0.590 | 0.710 |    |       |      |      |       |       |       |         | 0.780              | 8.6  | 36.0  | 36.0  |
| Kapoor and Picard [2005]          | Acc   | 0.668 |       |    | 0.820 |      |      |       |       | 0.572 |         | 0.865              | 5.5  | 29.5  | 51.2  |
| Karpouzis et al. [2007]           | Acc   | 0.670 | 0.730 |    |       |      |      |       |       |       |         | 0.820              | 12.3 | 22.4  | 22.4  |
| Kessous et al. [2010]             | Acc   | 0.483 | 0.571 |    | 0.671 |      |      |       |       |       |         | 0.783              | 16.7 | 37.1  | 62.1  |
| Khalali and Moradi [2009]         | Acc   |       |       |    |       |      |      |       |       |       |         | 0.622              | -6.7 | 20.3  | 20.3  |
| Kim et al. [2005]                 | Acc   |       | 0.520 |    |       |      |      |       | 0.517 | 0.667 |         | 0.660              | 24.5 | 26.9  | 26.9  |
| Kim [2007]                        | Acc   |       | 0.540 |    |       |      |      |       | 0.530 |       |         | 0.550              | 1.9  | 7.8   | 7.8   |
| Kim and Lingenfelter [2010]       | Acc   |       | 0.711 |    |       |      |      |       | 0.640 |       |         | 0.724              | 1.7  | 13.0  | 13.0  |
| Koelstra et al. [2012]            | F1    |       |       |    |       |      |      |       | 0.560 | 0.549 | 0.619   | 0.627              | 1.2  | 9.2   | 17.9  |
| Krell et al. [2013]               | Acc   | 0.553 | 0.605 |    |       |      |      |       |       |       |         | 0.798              | 31.9 | 44.4  | 44.4  |
| Lin et al. [2012]                 | Acc   | 0.713 | 0.710 |    |       |      |      |       |       |       |         | 0.906              | 27.0 | 27.6  | 27.6  |
| Lin et al. [2012]                 | Acc   | 0.621 | 0.603 |    |       |      |      |       |       |       |         | 0.781              | 25.7 | 29.5  | 29.5  |
| Lingenfelter et al. [2011]        | Acc   | 0.480 | 0.450 |    |       |      |      |       |       |       |         | 0.550              | 14.6 | 22.2  | 22.2  |
| Lingenfelter et al. [2011]        | Acc   | 0.530 | 0.610 |    |       |      |      |       |       |       |         | 0.610              | 0.0  | 15.1  | 15.1  |
| Litman and Forbes-Riley [2004]    | Acc   |       | 0.555 |    | 0.580 |      |      |       |       |       |         | 0.612              | 5.6  | 10.3  | 10.3  |
| Litman and Forbes-Riley [2006a]   | Acc   |       | 0.695 |    | 0.745 |      |      |       |       |       |         | 0.750              | 0.7  | 7.9   | 7.9   |
| Litman and Forbes-Riley [2006a]   | Acc   |       | 0.520 |    | 0.545 |      |      |       |       |       |         | 0.570              | 4.5  | 9.6   | 9.6   |
| Lu and Jia [2012]                 | Acc   | 0.622 | 0.760 |    |       |      |      |       |       |       |         | 0.911              | 20.0 | 46.8  | 46.8  |
| Mansoorizadeh and Charkari [2010] | Acc   | 0.540 | 0.510 |    |       |      |      |       |       |       |         | 0.770              | 42.6 | 51.0  | 51.0  |
| Mansoorizadeh and Charkari [2010] | Acc   | 0.370 | 0.330 |    |       |      |      |       |       |       |         | 0.710              | 91.9 | 115.2 | 115.2 |
| Metallinou et al. [2008]          | Acc   | 0.654 | 0.544 |    |       |      |      |       |       |       |         | 0.754              | 15.4 | 38.8  | 38.8  |
| Metallinou et al. [2012]          | Acc   | 0.562 | 0.559 |    |       |      |      |       |       |       |         | 0.630              | 2.5  | 24.5  | 24.5  |
| Monkaresi et al. [2012]           | F1    | 0.582 |       |    |       |      |      | 0.512 |       |       |         | 0.612              | 5.1  | 19.6  | 19.6  |
| Nicolaou et al. [2011]            | CC    | 0.603 | 0.515 |    | 0.502 |      |      |       |       |       |         | 0.719              | 10.7 | 32.4  | 67.8  |
| Pal et al. [2006]                 | Acc   | 0.640 | 0.742 |    |       |      |      |       |       |       |         | 0.752              | 1.3  | 17.5  | 17.5  |

Continued

## A Review and Meta-Analysis of Multimodal Affect Detection Systems

43:15

Table II. Continued

| Reference               | UM    |       | Text  | Body  | Gaze  | PPhy  | CPhy  | Content | MM    | MM Effect Size (%) |       |       |
|-------------------------|-------|-------|-------|-------|-------|-------|-------|---------|-------|--------------------|-------|-------|
|                         | Meas. | Face  |       |       |       |       |       |         |       | Voice              | MM1   | MM2   |
| Paleari et al. [2009]   | Acc   | 0.321 | 0.361 |       |       |       |       |         | 0.430 | 19.1               | 34.0  | 34.0  |
| Park et al. [2012]      | Acc   | 0.771 | 0.773 |       |       |       |       |         | 0.814 | 5.3                | 5.6   | 5.6   |
| Rabie et al. [2009]     | Acc   | 0.745 | 0.619 |       |       |       |       |         | 0.782 | 4.9                | 26.3  | 26.3  |
| Rashid et al. [2012]    | Acc   | 0.742 | 0.674 |       |       |       |       |         | 0.803 | 8.3                | 19.1  | 19.1  |
| Rigoll et al. [2005]    | Acc   |       | 0.742 | 0.596 |       |       |       |         | 0.920 | 24.0               | 54.4  | 54.4  |
| Rosas et al. [2013]     | Acc   | 0.610 | 0.468 | 0.649 |       |       |       |         | 0.750 | 15.5               | 22.9  | 60.4  |
| Rosas et al. [2013]     | Acc   | 0.540 | 0.486 | 0.540 |       |       |       |         | 0.649 | 20.0               | 20.0  | 33.3  |
| Rozgic et al. [2012]    | Acc   | 0.513 | 0.609 | 0.486 |       |       |       |         | 0.694 | 14.0               | 35.3  | 42.8  |
| Savran et al. [2012]    | CC    | 0.178 | 0.092 | 0.162 |       |       |       |         | 0.280 | 41.9               | 121.9 | 217.6 |
| Schuller et al. [2007]  | Acc   | 0.312 | 0.621 |       |       |       |       |         | 0.639 | 2.9                | 104.8 | 104.8 |
| Schuller [2011]         | CC    |       | 0.683 | 0.685 |       |       |       |         | 0.776 | 2.6                | 28.8  | 28.8  |
| Sebe et al. [2006]      | Acc   | 0.560 | 0.450 |       |       |       |       |         | 0.900 | 60.7               | 100.0 | 100.0 |
| Seppi et al. [2008]     | Acc   |       | 0.631 | 0.629 |       |       |       |         | 0.664 | 5.2                | 5.6   | 5.6   |
| Shan et al. [2007]      | Acc   | 0.792 |       |       | 0.726 |       |       |         | 0.885 | 11.7               | 21.9  | 21.9  |
| Soleymani et al. [2012] | Acc   |       |       |       |       | 0.689 |       |         | 0.725 | 5.2                | 29.3  | 29.3  |
| Tu and Yu [2012]        | Acc   | 0.600 | 0.570 |       |       |       | 0.563 |         | 0.720 | 20.0               | 26.3  | 26.3  |
| Vu et al. [2011]        | Acc   |       | 0.700 | 0.885 |       |       |       |         | 0.854 | -3.5               | 22.0  | 22.0  |
| Wagner et al. [2011]    | Acc   | 0.480 | 0.510 | 0.420 |       |       |       |         | 0.550 | 7.8                | 14.6  | 31.0  |
| Walter et al. [2011]    | Acc   |       | 0.764 |       |       | 0.722 |       |         | 0.778 | 1.8                | 7.8   | 7.8   |
| Wang and Guan [2005]    | Acc   | 0.493 | 0.664 |       |       |       |       |         | 0.700 | 5.4                | 42.0  | 42.0  |
| Wang and Guan [2008]    | Acc   | 0.493 | 0.664 |       |       |       |       |         | 0.700 | 5.4                | 42.0  | 42.0  |
| Wang et al. [2013]      | Acc   |       |       |       |       |       | 0.696 | 0.658   | 0.774 | 7.7                | 22.4  | 22.4  |
| Wimmer et al. [2008]    | Acc   | 0.611 | 0.737 |       |       |       |       |         | 0.818 | 11.0               | 33.9  | 33.9  |
| Wöllmer et al. [2010]   | Acc   | 0.497 | 0.511 |       |       |       |       |         | 0.672 | 21.6               | 48.3  | 48.3  |
| Wöllmer et al. [2013a]  | Acc   | 0.545 | 0.596 |       |       |       |       |         | 0.616 | 0.8                | 17.1  | 17.1  |
| Wöllmer et al. [2013b]  | Acc   | 0.612 | 0.644 | 0.730 |       |       |       |         | 0.720 | -1.4               | 11.8  | 17.6  |
| Wu and Liang [2011]     | Acc   |       | 0.800 | 0.809 |       |       |       |         | 0.836 | 3.3                | 4.4   | 4.4   |
| Zeng et al. [2005]      | Acc   | 0.390 | 0.690 |       |       |       |       |         | 0.750 | 8.7                | 92.3  | 92.3  |
| Zeng et al. [2006]      | Acc   | 0.862 | 0.701 |       |       |       |       |         | 0.899 | 4.3                | 28.2  | 28.2  |
| Zeng et al. [2007]      | Acc   | 0.386 | 0.664 |       |       |       |       |         | 0.724 | 9.0                | 87.6  | 87.6  |

Notes: Measure (Acc = Percent correct; CC = correlation coefficient; F1 = F measure), UM (PPhy = peripheral physiology; CPhy = central physiology; Content = Content/Context); MM = MM accuracy; MM Effect Size (MM1 and MM2 = percent MM improvement over best and second-best UM, respectively; MMMin = percent MM improvement over worst UM).

385 scores, or correlation coefficients to quantify performance.

$$\text{MM1 effect} = 100 * \frac{a_{12} - \max(a_1, a_2)}{\max(a_1, a_2)}. \quad (1)$$

386 In addition to the MM1 effect, MM2 and MMin effects were also computed as the  
 387 percent MM improvement over the second-best and worst UM detectors. These are  
 388 important metrics to test for inhibition effects, which occur when MM accuracies are  
 389 lower than underperforming UM detectors.

390 It is important to note three points about the data presented in Table II. First,  
 391 accuracy scores associated with the best-performing detector were used when *multiple*  
 392 *detectors* or *multiple fusion techniques* were considered for the *same* classification task.  
 393 For example, Soleymani et al. [2012] reported both feature-level and decision-level MM  
 394 accuracies. Decision-level fusion yielded higher accuracies, so only decision-level fusion  
 395 results were used in the subsequent analyses.

396 Second, several studies performed *multiple discriminations* on the same set of af-  
 397 fective states. For example, D'Mello and Graesser [2010] developed one classifier to  
 398 predict four affective states and another to predict an overlapping but different set of  
 399 five affective states. Similarly, the study by Eyben et al. [2011] contributed five data  
 400 points by independently predicting five affect dimensions (i.e., activation, expectancy,  
 401 intensity, power, and valence). In general, one data point was obtained for the studies  
 402 that performed a categorical classification. It was the dimensional studies that con-  
 403 tributed multiple data points because the number of models increases proportional  
 404 to number of dimensions considered. In all, data from 124 classification tasks was  
 405 obtained. These 124 data points were reduced to the 90 shown in Table II after the  
 406 aggregation procedure discussed next.

407 Third, when *multiple* classification tasks on the *same* dataset were performed, the one  
 408 closest to real-world performance was retained. For example, if text-based models were  
 409 built on automatically recognized and human-transcribed speech (e.g., Litman and  
 410 Forbes-Riley [2006b]), then the former was analyzed. Similarly, person-independent  
 411 validation results were used when both person-dependent and person-independent  
 412 validation methods were reported (e.g., D'Mello and Graesser [2010]). For the same  
 413 reason, event-level or segment-level analyses with a temporal resolution in seconds  
 414 were preferred over frame-level analyses with a temporal resolution in milliseconds  
 415 because affective phenomena operate across a coarser time span ranging from a few  
 416 seconds to tens of seconds [D'Mello and Graesser 2011; Rosenberg 1998].

#### 417 2.4. Data Treatment

418 Data from 124 classification tasks were subjected to aggregation, winsorization, and  
 419 standardization procedures as noted in the following.

420 *Aggregation.* Studies that performed multiple classification tasks on the same  
 421 dataset would bias the results and would violate independence assumptions of the  
 422 inferential statistical analyses applied to the data. Therefore, the data reported in Ta-  
 423 ble II consists of average scores across multiple classification tasks on the *same dataset*.  
 424 For example, the five correlation coefficients from the Eyben et al. [2011] study dis-  
 425 cussed previously were averaged to yield one data instance. Studies that reported  
 426 multiple classification tasks on *different datasets* were analyzed as separate data in-  
 427 stances (e.g., Rosas et al. [2013] where results corresponding to two distinct datasets  
 428 were reported in the same article).

429 *Winsorization (Outlier Treatment).* An examination of the MM, Max1, Max2, and  
 430 Min accuracy distributions did not yield any outliers, which, following standard con-  
 431 ventions, were defined as values exceeding three standard deviations from the mean.  
 432 However, the MM1, MM2, and MMin effects yielded two, one, and two outliers,



respectively. These outliers were replaced with the values corresponding to three standard deviations from the means of each distribution (60.7%→55.5%; 91.9%→55.5% for MM1 effect; 275%→168% for MM2 effect; and 217%→182% and 275%→182% for MM-Min effect), akin to a Winsorization procedure [Tukey and McLaughlin 1963], which is a widely used technique for outlier treatment. Paired-sample t-tests on the distributions before and after outlier replacement did not yield significant differences ( $p > 0.10$ ) for any of the three MM effects, thereby indicating that this method of treating outliers had no unintended effects.

*Standardization.* The three MM effects represent percent improvements over a baseline, so they are not sensitive to differences in accuracy metrics. However, raw detector accuracy scores were quantified in terms of percent correct (recognition accuracy), correlation coefficient, or F1 measure. These different metrics raised issues for the statistical methods used to analyze the raw detection accuracy scores (Max1, Max2, and Min). Hence, these measures were standardized (i.e., z-scores were computed) within each metric prior to the analyses.

### 3. RESULTS AND DISCUSSION

The results are presented with respect to the three major research questions listed in the Introduction: (a) What are the major trends in contemporary MM affect detectors? (b) What is the added improvement (if any) of MM affect detection accuracy (MM1 effects) over the best UM detectors? (c) Can we identify system-level factors identified in (a) that are predictive of MM1 effects analyzed in (b)?

It is useful to clarify our terminology before proceeding. *System* and *study* are used to refer to a multimodal affect detector (system) and its validation (study). *Effects* refer to percent improvement in MM accuracies over UM accuracies (MM1, MM2, and MMMin effects), while *accuracies* refer to affect detector performance represented as z-scores following metric-level standardization of percent correct, F1, and correlation coefficient (see Section 2.4).

#### 3.1. Major Trends in MM Affect Detectors

Table III lists descriptive statistics on the various system-level factors described in Section 2.2.

*Data Sources.* We note that on average MM detectors were constructed from affective data from 21.2 participants (not shown in Table III). There was also considerable variability ( $SD = 37.8$ ) in the number of participants used for model building, ranging from a single participant [Busso et al. 2004; Haq et al. 2008] to 343 participants [Wöllmer et al. 2013b]. An examination of the distribution indicated that 25% of the studies had five participants or fewer, 50% had 12 participants or fewer, and 97% of studies had fewer than 50 participants.

The data also indicated that the MM detectors were more likely to be trained on actor-portrayed affective displays (>50% of studies) rather than on more spontaneous expressions that were either experimentally induced or naturally occurred.

*Affect Models.* As is evident in Table III, approximately two thirds of the affect detectors focused on discrete (or categorical) affect models and performed classification tasks. Even though one third of the studies used dimensional models of affect, only 7.8% performed regressions. This was because several studies either collected categorical measures of affect dimensions (e.g., low or high arousal) or discretized continuous measures (e.g., via median splits or by applying clustering). On average, the classifiers discriminated 4.71 affective states ( $SD = 2.28$ ; median = 4 states), with a minimum of 2 and a maximum of 12 (not shown in Table III). The results also revealed that approximately one third of the affect detectors exclusively focused on discriminating the basic emotions, while less than 10% primarily focused on nonbasic emotions. Even

Table III. Descriptive Statistics on Study Features

| Dimension         | Prop. | Dimension         | Prop. |
|-------------------|-------|-------------------|-------|
| Data type         |       | Measure. model    |       |
| Acted             | 0.522 | Disc.             | 0.644 |
| Induced           | 0.278 | Dim.              | 0.356 |
| Natural           | 0.200 |                   |       |
| Detection model   |       | Affect detected   |       |
| Classification    | 0.922 | Disc. basic       | 0.367 |
| Regression        | 0.078 | Disc. nonbasic    | 0.078 |
|                   |       | Disc. mixed       | 0.178 |
| No. of modalities |       | Dim. simple       | 0.278 |
| Bimodal           | .867  | Dim. complex      | 0.100 |
| Trimodal          | .133  |                   |       |
|                   |       | Fusion method     |       |
| Modality          |       | Feature           | 0.389 |
| Face              | 0.767 | Decision          | 0.356 |
| Voice             | 0.822 | Hybrid            | 0.056 |
| Text              | 0.167 | Model             | 0.200 |
| Body              | 0.133 |                   |       |
| Eye Gaze          | 0.011 | Validation method |       |
| Peri. physio.     | 0.111 | Person indep.     | 0.378 |
| Central physio.   | 0.056 | Person dep.       | 0.622 |
| Content           | 0.067 |                   |       |

*Notes:* Prop. = Proportion; Peri = Peripheral; Physio. = Physiology; Content = Content/Context; Measure. = Measurement; Disc. = Discrete; Dim. = Dimensional; Indep. = Independent; Dep. = Dependent.

483 though 17.8% of the studies included a mixture of basic and nonbasic emotions, these  
 484 studies mainly focused on basic emotions with one or two nonbasic emotions. Hence,  
 485 more than 50% of the studies had a primary focus on the basic emotions.

486 The two primary dimensions of valence and arousal dominated the dimensional  
 487 models (approximately 30% of studies) with 10% of studies modeling more complex  
 488 dimensions. In all, 48 affective states (including dimensions) were modeled in the 90  
 489 studies (not shown in Table III). Only nine of the 48 affective states (18.8%) appeared  
 490 in more than 5% of the studies, and these nine states collectively accounted for 76% of  
 491 the states detected across all studies. The nine frequent states were (a) the six basic  
 492 emotions—anger (12%), sadness (11%), happiness (9%), fear (7%), disgust (7%), and  
 493 surprise (7%); (b) the two primary dimensions of valence (8%) and arousal (7%); and  
 494 (c) the state of no apparent feeling (8%) or neutral.

495 *Modalities.* The face and voice were the most commonly used modalities, each oc-  
 496 ccurring in over 75% of the studies. Text, body movements, and peripheral physiology  
 497 were individually used in at least 10% of the studies. Eye gaze, central physiology, and  
 498 context/content models were relatively infrequent.

499 Fifteen unique MM combinations were noted in the 90 studies. Of these, most were  
 500 bimodal (86.7%) systems, while a handful were trimodal systems. Audiovisual systems  
 501 (face + voice) comprised 55.6% of the MM systems, followed by speech + text (11.1%)  
 502 and face + speech + text (5.6%). These three combinations accounted for 72.3% of  
 503 the systems. In addition, voice + peripheral physiology, face + body movements, and  
 504 face + voice + body movements each accounted for 4.4% of the MM systems. In all,  
 505 these six MM combinations accounted for 85.6% of the systems, while the remaining  
 506 nine combinations were quite infrequent (each observed in <4% of the studies).

507 *Fusion Methods.* Several studies tested multiple fusion methods, so it was difficult  
 508 to accurately estimate if a particular method was used more frequently than others.

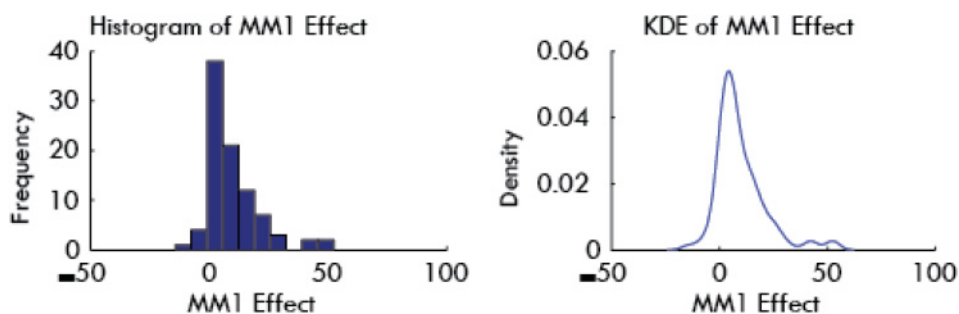


Fig. 1. Histogram (left) and kernel smoothing density estimation (right) of distribution of MM1 effects.

When multiple methods were used in the same study, we only recorded the method that yielded the best performance, because the final detector would presumably use the best-performing method. As noted in Table III, feature-level and decision-level fusion were dominant and were collectively observed in approximately 75% of the studies. Model-level fusion was somewhat less frequent (20%), but occurred at nontrivial rates. Data-level fusion was nonexistent and hybrid fusion was rare.

The most common feature-level fusion strategy simply involved concatenating feature vectors from individual modalities (e.g., D’Mello and Graesser [2010] and Forbes-Riley and Litman [2004]) with or without feature selection. The decision-level fusion methods usually relied on simple voting rules (e.g., Dy et al. [2010] and Gajsek et al. [2010]), but more nuanced ways of decision making were also proposed. Some of these include metadecision trees [Wu and Liang 2011], cascading specialists [Kim and Lingensfeler 2010; Wagner et al. 2011], Kalman filters [Glodek et al. 2013], Bayesian belief integration [Chanel et al. 2011], and Markov decision networks [Krell et al. 2013]. There was considerable variation in model-level fusion methods, but bidirectional long short-term memories [Eyben et al. 2010; Metallinou et al. 2012; Wöllmer et al. 2010, 2013a], various HMM-based approaches (error-weighted semicoupled HMMs [Lin et al. 2012], multistream HMMs [Zeng et al. 2005, 2007], boosted multistream HMMs [Zeng et al. 2006], boosted coupled HMMs [Lu and Jia 2012]) and Bayesian-based approaches (e.g., Jiang et al. [2011], Paleari et al. [2009], Sebe et al. [2006], and Wang et al. [2013]) were most prominent.

*Validation Methods.* Tenfold cross-validation at the segment (or frame) level was the most popular validation method. This method was used in 62.2% of the studies. This validation method is problematic when the goal is to build person-independent models (which is usually the goal), since instances from the same individual are in both the training and testing sets. In contrast, leave-one-subject-out or leave-several-subjects-out validation methods guarantee training and testing independence, but were used with considerably less frequency (37.8% of studies).

### 3.2. MM Effects and Accuracy

The data were analyzed in terms of (a) MM improvement over best UM accuracies (MM1 effects), (b) MM improvement over second-best (MM2 effects) and worst (MMMin effects) UM accuracies, and (c) relationships between UM and MM accuracies.

*Overall MM Effects (MM1 Effect).* The distribution of MM1 effects is presented in Figure 1. A one-sample t-test indicated that the mean MM1 effect of 9.83% significantly differed from zero,  $t(89) = 8.08$ ,  $p < 0.001$ ,  $d = 0.85$  sigma (large effect<sup>1</sup>). This suggests

<sup>1</sup>Cohen’s  $d$  is a common effect size statistic in standard deviation units (sigma) between two samples with means  $\bar{M}_1$  and  $\bar{M}_2$  and standard deviations  $\bar{s}_1$  and  $\bar{s}_2$  [Cohen 1992]. According to Cohen, effect

43:20

S. K. D'mello and J. Kory

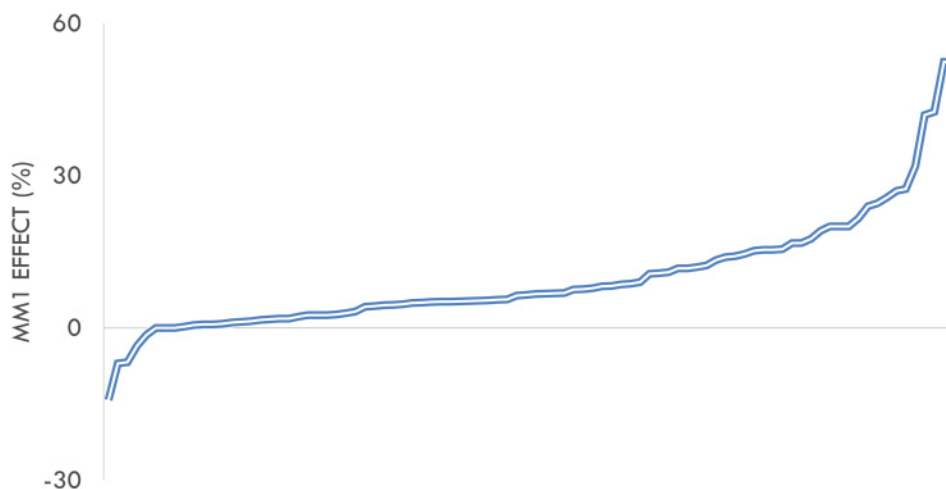


Fig. 2. MM1 effects (Y axis) by study number (X axis) ordered by effect size (ascending order).

Table IV. Grouping of MM1 Effects

| Group                     | Number of Studies | Percent of Studies (%) | Cumulative Percent (%) |
|---------------------------|-------------------|------------------------|------------------------|
| MM1 $\leq -1$             | 5                 | 5.56                   | 5.56                   |
| $-1 < \text{MM1} \leq 1$  | 8                 | 8.89                   | 14.4                   |
| $1 < \text{MM1} \leq 5$   | 21                | 23.3                   | 37.8                   |
| $5 < \text{MM1} \leq 10$  | 23                | 25.6                   | 63.3                   |
| $10 < \text{MM1} \leq 20$ | 20                | 22.2                   | 85.6                   |
| $20 < \text{MM1} \leq 30$ | 8                 | 8.89                   | 94.4                   |
| MM1 $> 30$                | 5                 | 5.56                   | 100.0                  |

Q1

544 that, on average, the MM detectors yield positive improvements in performance com-  
 545 pared to the best UM detectors.

546 There was considerable variance in the MM1 effect distribution. MM1 effects ranged  
 547 from  $-14.2\%$  to  $52.5\%$  with a standard deviation of  $11.5\%$ . The large range and the  
 548 fact that the standard deviation was greater than the mean, suggests that the *median*  
 549 *value* of  $6.60\%$  might provide a more accurate estimate of the central tendency of the  
 550 distribution than the mean.

551 To examine the distribution of MM1 effects more closely, we sorted the distribution  
 552 (see Figure 2), divided it into several categories of practical interest (see Table IV) and  
 553 computed the percent of studies falling into each category. This analysis indicated that  
 554 Q2 14.4% of the studies either yielded negative or negligible ( $\leq 1\%$ ) MM1 effects. Results  
 555 for the remaining 85% of the studies were much more positive in that roughly half  
 556 of the studies yielded either small 1%–5% or medium-sized (5%–10%) MM1 effects.  
 557 Approximately 35% of the studies yielded impressively large effects ( $> 10\%$ ).

558 *MM2 and MMMin Effects.* MM2 and MMMin effects are identical for the studies that  
 559 only considered two modalities (87% of studies), yet we analyze these effects separately  
 560 because there were some subtle differences in their distributions. MM2 effects ranged  
 561 from  $4.40\%$  to  $168.4\%$  with an impressive mean of  $40.0\%$  ( $SD = 36.9\%$ ). MMMin effects  
 562 had a mean of  $43.7\%$  ( $SD = 40.0\%$ ) and a range of  $4.40\%$ – $182.3\%$ . Given the large

sizes approximately equal to 0.3, 0.5, and 0.8 represent small, medium, and large effects, respectively.

$$d = (M_1 - M_2) / \sqrt{\frac{s_1^2 + s_2^2}{z}}$$

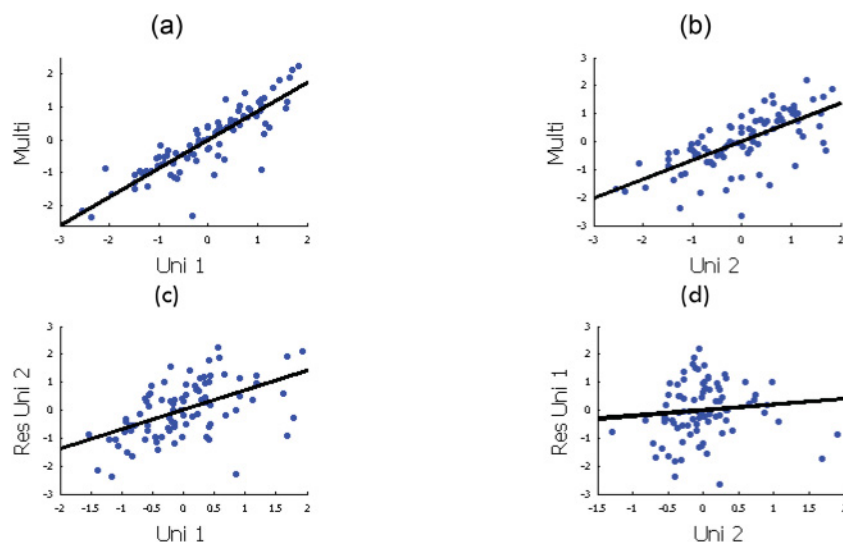


Fig. 3. Scatter plots denoting relationships between MM and UM accuracy along with regression line for (a) regression of MM (Multi) on best UM (Uni 1) accuracy; (b) regression of MM (Multi) on second-best UM (Uni 2) accuracy; (c) same as (a) but after controlling for second-best UM accuracy; and (d) same as (b) but after controlling for best UM accuracy.

standard deviations, the median values of 27.9% and 29.4% for MM2 and MMin effects, respectively, might be a more accurate summary statistics of these distributions. One-sample t-tests indicated that the mean MM2 effect significantly differed from zero,  $t(89) = 10.3$ ,  $p < 0.001$ ,  $d = 0.1.08$  sigma, as did the mean MMin effect,  $t(89) = 10.4$ ,  $p < 0.001$ ,  $d = 1.09$  sigma. Furthermore, paired samples t-tests indicated that the mean MM2 effect was significantly,  $t(89) = 8.18$ ,  $p < 0.001$ , and substantially ( $d = 1.11$  sigma) greater than the mean MM1 effect (9.83%). A similar finding was discovered when MMin effects were compared to MM1 effects,  $t(89) = 8.59$ ,  $p < 0.001$ ,  $d = 1.15$  sigma. In general, MM2 and MMin effects were approximately four times greater than MM1 effects, so MM detectors were substantially more accurate than their less effective UM counterparts.

*Relationships between UM and MM Accuracies.* There was a very robust correlation between best UM and MM accuracies,  $r(88) = 0.870$ ,  $p < 0.001$ . The correlation between second-best UM and MM accuracies was notable, but smaller,  $r(88) = 0.681$ . Best and second-best UM accuracies were also strongly correlated,  $r(88) = 0.725$ ,  $p < 0.001$ .

We simultaneously regressed MM accuracy (dependent or predicted variable) on best and second-best UM accuracies (independent or predictor variables). The model was significant,  $F(2, 87) = 139.7$ ,  $p < 0.001$ , and explained a robust amount of the variance,<sup>2</sup>  $R^2 = 0.763$ ;  $f^2 = 3.22$ . The best UM accuracy was a significant predictor ( $\beta = 0.795$ ,  $p < 0.001$ ) but second-best UM accuracy was not ( $\beta = 0.104$ ,  $p = 0.174$ ). This indicates that much of the variance in MM accuracy can be explained by the best UM accuracy.

These patterns are shown in Figure 3, where we note that the linear relationship between MM and best UM accuracy (Figure 3(a)) is retained after controlling for second-best UM accuracy (Figure 3(c)). However, the linear relationship between MM and

<sup>2</sup> $R^2$  or the coefficient of determination is used to assess goodness of fits of regression models. Using Cohen's recommended conventions [Cohen 1992], effect sizes are expressed as Cohen's  $f^2 = \frac{R^2}{1-R^2}$  and values of 0.02, 0.15, and 0.35 are taken to signify small, medium, and large effects, respectively.

588 second-best UM accuracy (Figure 3(b)) essentially disappears after controlling for best  
 589 UM accuracy (flat line in Figure 3(d)).

590 Hence, the final model simply consisted of predicting MM accuracy from best UM  
 591 accuracy. This model was significant,  $F(1, 88) = 274.8$ ,  $p < 0.001$ , and robust,  $R^2 =$   
 592  $0.757$ ,  $f^2 = 3.12$ . The standardized model coefficient ( $\beta$  weight) was 0.870, which  
 593 indicates that a 1 unit (in standard deviation units) increase in best UM accuracy  
 594 results in a 0.870 unit increase in MM accuracy.

595 To address the question of whether this regression model generalizes to new studies,  
 596 we performed a between-study 10-fold cross-validation analysis, which yielded an  $R^2$   
 597 of 0.746, which was very similar to  $R^2$  on the entire training set (0.757). The very small  
 598 discrepancy of 0.011 suggests that the regression model is expected to generalize to  
 599 new studies.

600 There is the question of whether MM accuracy increases, decreases, or remains  
 601 unchanged as a function of the difference between best and second-best UM accuracies.  
 602 To address this question, we retained the residuals (prediction errors or unexplained  
 603 variance) after regressing best on second-best UM accuracies. MM accuracy was then  
 604 regressed on the residual. The resultant model was significant and explained a modest  
 605 amount of variance,  $F(1, 88) = 37.6$ ,  $p < 0.001$ ,  $R^2 = 0.299$ ,  $f^2 = 0.43$ ,  $\beta = 0.574$ .  
 606 This finding suggests that MM accuracy improves in relation to the difference between  
 607 best and second-best UM accuracies. Put simply, MM accuracy was higher when UM  
 608 accuracies were more independent.

### 609 3.3. Moderation Analysis

610 Section 3.1 analyzed general trends in the design of MM affect detectors (system-level  
 611 factors) while Section 3.2 quantified performance in terms of MM effects. In this section,  
 612 we assess whether the system-level factors can predict MM performance.

613 The analyses proceeded by independently regressing MM1 effects and MM accuracy  
 614 on the eight system-level factors listed in Table III plus the number of participants  
 615 and number of affective states (10 total). Eight out of these 10 factors were categor-  
 616 ical variables, so these were dummy coded prior to constructing the models. It was  
 617 not possible to consider every unique modality combination given that there were 15  
 618 modality combinations and only 90 data points. However, since 55.6% of the modality  
 619 combinations were face + voice, we created a new indicator variable and coded it as a 1  
 620 for *face + voice* and a 0 for *other* modality combinations. Furthermore, given that only  
 621 five studies reported hybrid fusion, these studies were removed prior to constructing  
 622 the model for fusion method.

623 *Predicting MM1 Effects.* The resultant models for predicting MM1 effects are shown  
 624 in Table V, where  $k$  is the number of studies used to construct each model.  $F$  is the  
 625 test statistic for model significance ( $p$  value is in parentheses) and  $R^2$  is the measure  
 626 of model fit. Significant ( $p < 0.05$ ) models were discovered for data type, number of  
 627 affective states, and classifier fusion method, but not for the remaining seven factors.

628 The significant model for data type yielded a small- to medium-sized effect ( $f^2 =$   
 629  $0.087$ ). A test of model coefficients indicated that MM1 effects for detectors built from  
 630 natural data were statistically equivalent to those built from induced data ( $p = 0.299$ ),  
 631 but were significantly ( $p = 0.009$ ) lower than detectors built from acted data. The  
 632 induced models yielded quantitatively lower MM1 effects than the acted models, but  
 633 the difference was not quite significant ( $p = 0.102$ ). These patterns are graphically  
 634 depicted in Figure 4(a), where we note a negative linear relationship between MM1  
 635 effects and authenticity of training and validation data (mean MM1 effects = 12.7%,  
 636 8.19%, and 4.59% for acted, induced, and natural data, respectively). More precisely,  
 637 if data type is numerically coded along an authenticity dimension, with 1, 2, and 3

Table V. Regression Models for Predicting MM1 Effects

| Dimension                         | $k$ | Significance and Fit |       |
|-----------------------------------|-----|----------------------|-------|
|                                   |     | $F(p)$               | $R^2$ |
| Number of participants            | 88  | 0.004 (0.947)        | 0.000 |
| Data type                         | 90  | **3.80 (0.026)       | 0.080 |
| Affect representation model       | 90  | 1.25 (0.267)         | 0.014 |
| Affect detection model            | 90  | 0.329 (0.567)        | 0.004 |
| Affect states detected            | 90  | 0.828 (0.511)        | 0.037 |
| Number of affective states        | 83  | **6.77 (0.011)       | 0.077 |
| Number of modalities              | 90  | 1.02 (.316)          | .011  |
| Modality (face + voice vs. other) | 90  | 2.08 (.153)          | .023  |
| Fusion method                     | 85  | **4.96 (0.009)       | 0.108 |
| Validation method                 | 90  | 0.133 (0.716)        | 0.002 |

Note: \*\*denotes significant models at the  $p < 0.05$  level.

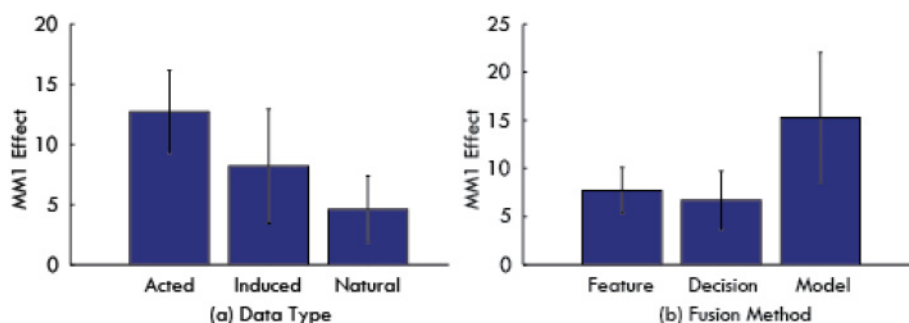


Fig. 4. Mean MM1 effect by (a) data type and (b) fusion method. Error bars are 95% confidence intervals.

representing acted, induced, and natural data, respectively, then there is a negative  $-0.245$  ( $p = 0.020$ ) correlation between data authenticity and MM1 effects.

The results also indicated that MM1 effects could be predicted from the number of affective states in the 85 studies that built classifiers instead of regressors. This model also yielded a small- to medium-sized effect ( $f^2 = 0.083$ ). Interestingly, the number of affective states was a positive predictor ( $\beta = 0.278$ ), so MM1 effects improved when more affective states were considered. One tentative interpretation of this finding is that the classification problem becomes more difficult when more affective states are considered and the additional modalities have more to contribute in this situation.

The third significant model had MM fusion type as the predictor and also yielded with a small- to medium-sized effect ( $f^2 = 0.121$ ). An analysis of the model coefficients indicated that MM1 effects associated with feature- ( $M = 7.73\%$ ) and decision-level ( $M = 6.68\%$ ) fusion were statistically equivalent ( $p = 0.661$ ), but were lower than MM1 effects for model-based fusion ( $M = 15.3\%$ ,  $p < .05$ ; see Figure 4(b)). This finding should be interpreted with caution because it *does not* represent direct comparisons of different fusion techniques on the *same* datasets and classification tasks. Instead, it simply suggests that, on average, model-level fusion yielded higher MM1 effects than feature-level and decision-level fusion.

*Predicting MM Accuracy.* In Section 3.2, we reported that 75.7% of the variance in MM accuracy was explained by the best UM accuracy. We investigated if this model could be improved by adding system-level factors. The analyses proceeded by testing if each system-level factor explained unique variance in MM accuracy after accounting for best UM accuracy (our previous model). This was accomplished with 10 hierarchical linear regressions with UM accuracy as the predictor for the Step 1 models and each

662 system-level factor as individual predictors in the Step 2 models. A significant change  
 663 in  $R^2$  from Step 1 to Step 2 would indicate that the system-level feature under con-  
 664 sideration explained additional variance in MM accuracy above and beyond best UM  
 665 accuracy.

666 The results yielded significant  $R^2$  changes ( $\Delta R^2$ ) for data type ( $\Delta R^2 = 0.034$ ,  $p =$   
 667  $0.002$ ), affect representation model ( $\Delta R^2 = 0.011$ ,  $p = 0.046$ ), number of affective  
 668 states classified ( $\Delta R^2 = 0.025$ ,  $p = 0.005$ ), and fusion method ( $\Delta R^2 = 0.014$ ,  $p =$   
 669  $0.041$ ), but not for number of subjects ( $\Delta R^2 = 0.001$ ,  $p = 0.633$ ), affect detection model  
 670 ( $\Delta R^2 = 0.00$ ,  $p = 1.00$ ), affect states detected ( $\Delta R^2 = 0.019$ ,  $p = 0.144$ ), number of  
 671 modalities ( $\Delta R^2 = 0.00$ ,  $p = 0.936$ ), modality (face + voice vs. other)—( $\Delta R^2 = 0.009$ ,  
 672  $p = 0.068$ ), and validation method ( $\Delta R^2 = 0.06$ ,  $p = 0.137$ ).

673 Examining coefficients of models with significant  $\Delta R^2$  indicated that (a) detectors  
 674 developed from induced and natural affect had MM accuracies that were on par but  
 675 significantly ( $p < 0.01$ ) lower than detectors developed from acted data, (b) detectors  
 676 that used discrete affect models yielded significantly ( $p = 0.043$ ) higher accuracies  
 677 than their dimensional counterparts, (c) MM accuracies increased ( $p = 0.005$ ) when  
 678 more affective states were classified, and (d) model-level fusion resulted in significantly  
 679 ( $p < 0.05$ ) higher MM accuracies than feature- and decision-level fusion.

680 Next, we created a model that predicted MM accuracy when these four key factors  
 681 (data type, affect representation model, number of affective states, and fusion method)  
 682 were considered simultaneously. This model was constructed using a *forward feature*  
 683 *selection* approach, where features were incrementally added if they improved model  
 684 fit. It should be noted that due to missing data (elimination of five studies that used  
 685 hybrid fusion and number of states not applicable in the seven studies that developed  
 686 regressors), this model was constructed from 78 out of the 90 studies. The Step 1 model  
 687 on these 78 studies with the best UM accuracy as a predictor yielded an  $R^2$  of 0.796  
 688 (note the difference from the 0.757  $R^2$  reported earlier on all 90 studies). The Step 2  
 689 model had an  $R^2$  of 0.832, which represented a significant improvement ( $\Delta R^2 = 0.036$ ,  
 690  $p = 0.014$ ) from the Step 1 model. The significant predictors that were retained by  
 691 forward feature selection were best UM accuracy ( $\beta = 0.879$ ,  $p < 0.001$ ), whether the  
 692 training data was acted (coded as 1) or not (coded as 0) ( $\beta = 0.138$ ,  $p = 0.006$ ), and  
 693 whether model-level fusion (coded as 1) was used in lieu of feature and decision fusion  
 694 (coded as 0) ( $\beta = 0.122$ ,  $p = 0.014$ ). Finally, 10-fold cross-validation yielded an  $R^2$  of  
 695 0.803. The very small discrepancy of 0.029 from  $R^2$  on entire training data is suggestive  
 696 of excellent generalizability of the final model.

## 697 4. GENERAL DISCUSSION

698 Timely surveys that synthesize research are critical in any burgeoning research area.  
 699 The qualitative nature of surveys can be complemented with quantitative meta-  
 700 analyses, an invaluable scientific tool for approximating a population variable from  
 701 effects obtained in individual studies that vary along multiple dimensions [Borenstein  
 702 et al. 2009]. In this article, we identified 90 contemporary MM affect detectors from  
 703 the peer-reviewed literature, coded and descriptively analyzed each detector along 10  
 704 dimensions, performed a meta-analysis on MM accuracy as compared to UM accuracy  
 705 (MM effects), and identified important system-level moderators of MM1 effects. In  
 706 this section, we summarize our major findings along with their applied implications,  
 707 discuss their theoretical implications, address limitations, offer recommendations for  
 708 future work, and make concluding remarks.

### 709 4.1. Major Findings and Applied Implications

710 The major findings are organized with respect to the three research questions listed  
 711 in the Introduction: (a) identifying major trends in MM affect detectors, (b) analyzing



MM effects and MM accuracy, and (c) identifying the factors that moderate MM effects and accuracies. 712

*Major Trends in MM Affect Detectors.* The first surveys on automated affect detection emerged over a decade ago [Cowie et al. 2001; Pantic and Rothkrantz 2003]. According to these pioneering surveys, and at the risk of overgeneralization, the state of the art in affect detection in 2003 and earlier could be summarized 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.” Based on the present analysis, subjective interpretation, and somewhat overgeneralization, the 2013 state of the art can be summarized as “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.” The italicized items in the previous summary reflect important changes in the state of the art from 2003 to 2013. Based on this comparison, it is clear that considerable progress has been made, although there is still more to be done. We discuss some of the remaining issues with respect to the following four aspects: authenticity, utility, scope, and generalizability. 713

*Authenticity* refers to the naturalness of training and validation data and is directly related to the extent to which an affect detector developed in the lab can be applied in the real world. The fact that more than 50% of the affect detectors were based on acted data is of some concern since spontaneous and acted expressions differ in surprising ways. A striking example is a study that found that individuals rarely smile when generating posed expressions of frustration, but smiles were discovered in 90% of instances of spontaneous frustration [Hoque and Picard 2011]. 714

*Utility* refers to whether the affect detectors can be expected to be useful in real-world contexts. Assuming that detection accuracy will eventually be sufficiently accurate, the question is whether the affective states that are detected are relevant in the real-world contexts of use (e.g., editing a word document on a computer). This is a critical issue since more than 50% of the studies primarily focused on detecting the basic emotions of anger, sadness, fear, frustration, disgust, and surprise. This is a bit unfortunate because it has been asserted that many interactions with computers and even human-human interpersonal communication rarely involve the basic emotions [Cowie et al. 2005; Zeng et al. 2009]. Some recent evidence for this assertion can be found in a meta-analysis on 24 studies that collectively tracked the emotions of over 1,700 students during interactions with a range of learning technologies [D’Mello 2013]. The major finding was that engagement, confusion, boredom, curiosity, frustration, and happiness were the most frequent affective states. With the exception of happiness, which occurred with some frequency, the basic emotions were rarely observed in over 1,200 hours of interaction. 715

*Scope* (in this context) simply refers to the landscape of configurations that were covered by the affect detectors. In addition to the basic versus nonbasic emotion imbalance discussed previously, perhaps the greatest disparity emerges in the modality combinations. More specifically, the eight modalities identified in Table III afford 28 and 56 unique bimodal and trimodal combinations, respectively. However, only 15 out of the possible 84 (28 + 56) combinations (17.9%) were observed at least once in the data. Six of these (7.14% of possible combinations) were represented in more than 85% of the studies, while the face + voice, which represents a mere 1.19% of possible modality combinations, was the focus of more than half of the studies. Indeed, the explored 716

764 MM space is sparse and there is both the room for and the need to consider different  
765 modality combinations.

766 *Generalizability* pertains to an affect detector's ability to maintain its level of ac-  
767 curacy when applied to new individuals and to new or related contexts. One way to  
768 facilitate generalizability is to collect training data in diverse contexts and from a large  
769 number of individuals. There is clearly more work to be done in this respect since 97%  
770 of the studies collected training and validation data from fewer than 50 individuals and  
771 usually in a single context (e.g., watching videos, interacting with a specific interface).  
772 Generalizability across the individual can be assessed via person-independent models,  
773 where training and validation data are completely independent. As noted in Table III,  
774 about 40% of the studies used person-independent validation methods, so there is some  
775 confidence on their generalizability (across individuals). Unfortunately, no clear case  
776 for generalizability can be made for the remaining 60% of studies that used person-  
777 dependent validation methods. Furthermore, no notable efforts were made to assess  
778 generalizability across tasks, situational contexts, datasets, and cultures. This is par-  
779 ticularly important since emerging data suggests that models trained on individuals  
780 from one demographic do not necessarily generalize to another [Ocumpaugh et al.  
781 2014].

782 *MM Effects and Accuracy.* A number of important conclusions can be drawn from the  
783 analysis of MM effects and MM and UM accuracies. Over 85% of the studies resulted  
784 in MM1 effects greater than at least 1%. This provides important evidence that MM  
785 classifiers do outperform their best UM counterparts. The sizes of the mean (9.83%)  
786 and median (6.60%) MM1 effects resemble modest improvements over UM accuracy.  
787 Importantly, however, MM1 effects associated with detectors trained on naturalistic  
788 data (4.59%) were three times lower than detectors trained on acted data (12.7%).  
789 Since the ultimate goal of affect detection is to sense naturalistic affective expressions,  
790 the modest 4.59% effect might represent a more accurate estimate of state-of-the-art  
791 multimodal affect detection improvement.

792 The question of whether this modest improvement in accuracy obtained by MM  
793 systems is worth their increased complexity is a question that is best addressed at  
794 the application level. It should also be noted that the present study only evaluated  
795 MM detectors from a single dimension, namely, performance improvements over UM  
796 detectors. However, MM detectors have additional advantages, such as providing higher  
797 fidelity models of affect expression and the ability to address missing data problems  
798 that can cripple UM detectors. Furthermore, the analysis that focused on assessing MM  
799 performance improvements over the second-best and worst UM classifier indicated  
800 that although combining modalities yields modest improvements in affect detection  
801 accuracies, *considering multiple individual modalities* can have a major impact on  
802 performance. This is because performance would be severely impacted if only one  
803 modality was modeled and in the worst case if it always happened to be the lower  
804 performing modality.

805 Turning back to MM1 effects, one reason for their relatively modest size, especially  
806 for the systems trained on more naturalistic data, is that there might be consider-  
807 able redundancy among the different modalities. Strong correlations among the best  
808 UM, second-best UM, and MM accuracies provide some evidence to support this view.  
809 Evidence for redundancy among modalities can also be obtained by the fact that the  
810 best UM accuracies predicted 75.7% of the variance in MM accuracies and this finding  
811 generalizes to new studies. Impressive MM1 effects are not expected if the different  
812 modalities convey similar information, albeit in different ways. The analysis that found  
813 that MM accuracies increased when UM accuracies were more dissimilar provides some  
814 evidence in support of this claim.

The lower multimodal effects for natural emotional expressions compared to acted expressions might also be attributable to several differences among the two. In particular, some aspects of acted expressions that are conducive to multimodal effects include increased intensity (since they are usually exaggerated), decreased variability (since they are generated out of context), increased coordination between different modalities (since prototypical emotions are invoked), and increased specificity (since there is lower likelihood of multiple emotions being experienced) [Barrett 2006; Russell 2003].

*Factors that Moderate MM Effects.* We examined 10 system-level factors and identified three that moderated MM1 effects. We discovered that MM1 effects were positively impacted by acted data (vs. induced or natural data), number of affective states classified, and when model-level modality fusion methods were used (vs. feature or decision level). Two out of these four system-level factors (acted vs. nonacted data and model-level vs. non-model-level fusion) yielded a 3.6% improvement in predicting MM accuracy over best UM accuracy. Furthermore, fit of the final model with all three predictors was excellent ( $R^2$  of 0.832), and generalizes to new studies as verified with a 10-fold study-level cross-validation analysis.

The final model, specified in Equation (2), can be used by researchers to predict expected multimodal classification accuracy (proportion of cases correctly classified ranging from 0 to 1) *prior* to even constructing the classifiers. *Best unimodal accuracy* is the classification accuracy (as a proportion ranging from 0 to 1) of the best UM detector. *Data type acted* is an indicator variable set to 1 for acted data and 0 for induced data. *Model-level fusion* is also an indicator variable set to 1 for model-level fusion and 0 for feature- and decision-level fusion.

$$\begin{aligned} \text{MM accuracy} = & 0.900 \times \text{Best unimodal accuracy} + 0.273 \times \text{Data type acted} \\ & + 0.312 \times \text{Model level fusion} - 0.253 \end{aligned} \quad (2)$$

#### 4.2. Theoretical Implications

The fact that combining MM accuracies yielded modest improvements has important implications for psychological theories of emotion. These theories in turn guide much of the affect detection models, so alignment of our findings with emotion theory has implications for next-generation affect detection systems.

The classical model of emotion, which was proposed by Tomkins [1962], Ekman [1992], and Izard [2007], and others, posits that discrete “affect programs” produce the physiological, behavioral, and subjective changes associated with a particular emotion. According to this theory of “basic emotions,” there is a specialized circuit for each basic emotion in the brain. Upon activation, this circuit triggers a host of *coordinated responses* in the mind and body. In other words, an emotion is expressed via a sophisticated synchronized response that incorporates peripheral physiology, facial expression, speech, modulations of posture, affective speech, and instrumental action. This prediction is very relevant to affect detection because it suggests that MM affect detection should be more reliable due to this coordinated recruitment of response systems.

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]. According to this contemporary view, there is no central affect program that coordinates the various components of an emotional episode. Instead, these components are loosely coupled and the specific context and appraisals determine which bodily systems are activated. These models would accommodate the prediction that in most cases a combination of modalities might conceivably yield small improvements in classification accuracies. Hence, other than the rare cases of prototypical emotions, or in artificial experimental contexts involving

862 acted emotions, modest multimodal effects might be expected. Indeed, this is exactly  
863 what was observed in the present analysis.

#### 864 4.3. Limitations and Future Work

865 There are five primary limitations to this work. The first pertains to the comprehen-  
866 siveness of the studies that were analyzed. Our goal was to obtain a reasonably large  
867 sample of MM studies rather than attempting to analyze every single study in the lit-  
868 erature. This is defensible because one does not need to study an entire population to  
869 estimate its parameters. Furthermore, almost all of the tests of statistical significance  
870 yielded significant results and we show evidence for model generalizability, thereby  
871 suggesting that our sample size of 90 studies was adequate to detect the relatively  
872 large effects in our data.

873 The second limitation was that there was some imbalance with respect to the modal-  
874 ities, data, evaluation metrics, and affective states classified. For example, a majority  
875 of the studies we analyzed focused on audio-visual affect recognition, so the results  
876 are somewhat biased toward these systems. It is important to note, however, that this  
877 imbalance in our study is linked to a similar imbalance in the current state of the art.  
878 Specifically, most studies focus on the audio and visual modalities, while central  
879 physiology, gaze, and content/context-based sensing are comparatively rare. Peripheral  
880 physiological-based affect sensing (i.e., biosignals) are quite common affect detection  
881 modalities, but these are not often combined with face, voice, text, and other modalities.

882 A third limitation that befalls all meta-analyses is the possibility of publication bias.  
883 This is because it is likely that the papers that report positive MM1 effects are more  
884 likely to be published, and subsequently included in this meta-analysis, than papers  
885 that report negligible or negative effects. We suspect that this might not be a severe  
886 issue in the present study, since approximately 15% of the studies reported negative or  
887 null (<1%) MM1 effects, but there is no clear way to assess publication bias with the  
888 present data.

889 A fourth limitation is that the present study is more consistent with an informal  
890 meta-analytic approach rather than a more formal meta-analysis procedure. This was  
891 due to a lack of available information needed to perform a formal meta-analysis. More  
892 specifically, one of the key steps in conducting a meta-analysis is to inversely weight the  
893 effect size with respect to its error, but error estimates on affect detection accuracies  
894 were never reported in the papers we analyzed. This also precluded the use of well-  
895 established techniques to identify and correct for publication bias like trim-and-fill  
896 procedures [Duval and Tweedie 2000].

897 Fifth, the somewhat large timespan (roughly 10 years) of the studies included in this  
898 analysis might also be of some concern since the newer classification and fusion meth-  
899 ods were unavailable for some of the older studies. Although the selection procedure did  
900 bias newer studies in lieu of older ones, it is possible that the older studies might have  
901 yielded better multimodal accuracies if some of the latest multimodal fusion methods  
902 were used. However, this does not appear to be a major concern as publication date  
903 (normalized so that the earliest study in 2004 was coded as 0, 2005 as 1, and so on) was  
904 not correlated with MM1 ( $r(88) = 0.042$ ,  $p = 0.696$ ), MM2 ( $r(88) = 0.056$ ,  $p = 0.600$ ),  
905 or MMMin ( $r(88) = 0.102$ ,  $p = 0.338$ ) effects. Nevertheless, it would be informative  
906 to reanalyze some of the older datasets with newer methods to ascertain if the use of  
907 newer techniques results in performance improvements.

#### 908 4.4. Recommendations for Future Systems

909 In this section we list some guidelines based on our analysis of the 90 multimodal  
910 detectors. These should be considered to be general recommendations since decisions  
911 should ultimately be guided by specific application contexts. Some of these suggestions

might seem obvious; however, they are noted here since some or all were ignored in more or less all of the studies. 912 913

First, there is a tradeoff between accuracy and authenticity in that highly accurate results are usually obtained in nonauthentic contexts, specifically building person-dependent models to detect acted expressions recorded in ideal conditions. Lower accuracies obtained in more naturalistic contexts are of greater practical value. Second, excellent results without meaningful comparison conditions are of less importance than modest results with stringent comparisons. For example, if a new multimodal fusion technique is being proposed, then its improvement over simpler techniques (e.g., naïve feature-level fusion) should be reported. Similarly, classification accuracy (or recognition rate) is a meaningless metric without a baseline comparison when there is an uneven distribution of classes (more on this point follows). Third, only a small subset of the landscape encompassing modalities and affective states has been explored. In addition to refining systems that operate on already-explored areas of this landscape, systems that explore new areas could lead to exciting innovations and discoveries. One suggestion is to focus on different modalities in addition to or in lieu of the face and speech to detect nonbasic affective states that pervade human-computer interactions, such as confusion, frustration, and perhaps even boredom. Fourth, model-level fusion techniques that embrace, rather than ignore, time-varying relationships among different modalities showed significant promise, so it might be useful to channel research efforts into improving these techniques. Fifth, the standard procedure of collecting labeled data to train supervised classifiers is inherently limited due to the manual affect annotation process, thereby resulting in small datasets (in terms of number of unique individuals). It is unlikely that this approach will lead to models that generalize at large [Ocuppaugh et al. 2014]; hence, it might be useful to consider semisupervised learning approaches that only require a small subset of the training data to be annotated. Furthermore, crowd-sourcing techniques might be useful alternatives to current cumbersome annotation methods that simply do not scale to larger datasets [McDuff et al. 2012]. 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940

It would also be highly beneficial if there was a more or less standard approach to evaluating and reporting results of affect detectors. Some suggested evaluation criteria include (a) meaningful comparison conditions when new systems are being proposed (as noted previously), (b) using person-independent validation techniques, (c) testing promising affect detectors developed by other researchers on one's own datasets (this was very rare), (d) testing new techniques on multiple datasets (i.e., cross-corpus evaluations), and (e) studying generalizability to individuals of different demographics—also referred to as population validity. 941 942 943 944 945 946 947 948

Suggestions on how to report results include reporting of (a) accuracy metrics that correct for uneven distribution of classes, (b) error estimates on accuracy measures, (c) number of individuals and instances, and (d) other information noted in Table III. With respect to the first item in this list, Jeni et al. [2013] recently evaluated a number of classification accuracy metrics by performing simulations as well as analyzing real datasets with imbalanced class distributions (skewed data). Their findings indicated that several of the commonly used metrics, such as accuracy (recognition rate), kappa, F-score, Krippendorff's alpha, and area under the precision-recall curve, were adversely affected by data skew. Area under the Receiver Operating Characteristics (ROC) curve (AUC or  $A'$ ) was most robust to data skew, but tended to minimize poor performance when compared to precision-recall curves. They recommended reporting both original uncorrected performance metrics as well as skew-normalized versions of these metrics with the normalization conducted by up-sampling and down-sampling the *test* partitions (the paper also provides a link to software to compute the skew-normalized statistics). 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963

964 **4.5. Concluding Remarks**

965 The phrase “*consistent, but modest under natural conditions*” succinctly captures per-  
966 formance of contemporary affect detectors. These MM detectors were *consistently* better  
967 than their UM counterparts, but the improvements were *modest* when the detectors  
968 were trained on naturalistic affect expressions. A fundamental question is whether  
969 these findings can be best explained by the *method* or by the *data*. In particular, were  
970 MM1 effects modest because the detectors are not sufficiently sophisticated to model  
971 the intricate nonlinear time-varied relationships between the different modalities? Or  
972 were they modest because the training data did not contain adequate expressions of  
973 coordination among modalities, thereby rendering even the most sophisticated detec-  
974 tors inept? The field of MM affect detection is too young to currently settle these issues,  
975 so an answer awaits further research.

976 However, there is another possibility beyond the method and the data. It may be the  
977 case that the expression of naturalistic emotions is inherently a diffuse phenomenon,  
978 which will yield modest effects irrespective of method or data. This suggests that in ad-  
979 dition to considering different methods and data sources, it might be useful to consider  
980 alternate models of emotion beyond the classic view described in Section 4.2. Thus  
981 far, the emphasis has been on the method and the data, at the expense of examining  
982 the affective phenomenon itself (i.e., insufficient attention to recent development in  
983 emotion theories and alternate models). Perhaps a more balanced approach that com-  
984 bines better data sources and innovative algorithms with more diverse emotion models  
985 represents the most promising way forward.

986 Whatever the case may be, this review and analysis has shown that the field of  
987 multimodal affect detection has come a long way from the initial proof-of-concept sys-  
988 tems of the past. Skeptics who thought that computers could never sense anything as  
989 elusive as affect have repeatedly been proven wrong. Even more significant is the fact  
990 that emerging systems go beyond detecting affect by dynamically responding to the  
991 sensed affect, thereby closing the so-called *affective loop* [Conati et al. 2005]. For exam-  
992 ple, the Affective AutoTutor is an intelligent tutoring system that improves learning  
993 gains for low domain-knowledge students by automatically sensing (via a MM anal-  
994 ysis of contextual cues, facial features, and body movements) and responding to con-  
995 fusion, frustration, and boredom [D’Mello and Graesser 2012]. UNC-ITSPPOKE is a  
996 speech-enabled intelligent tutoring system that automatically senses and responds to  
997 a learner’s uncertainty by modeling acoustic-prosodic and lexical features of students’  
998 spoken responses [Forbes-Riley and Litman 2011]. Another example is the Affective  
999 Music Player, which strategically selects music to induce specific moods (positive, neg-  
1000 ative, neutral) on a personalized basis via a predictive psychophysiological model [van  
1001 der Zwaag et al. 2013]. In general, systems that both sense and respond to affect are  
1002 continually emerging as documented in a recent edited volume on affective computing  
1003 [Calvo et al. 2014].

1004 Despite impressive progress, one limitation of most (but not all of these systems)  
1005 is that they have been tested in lab-based contexts (the Affective Music player is an  
1006 exception). Hence, the challenge is now to repudiate critics who think that affective  
1007 systems will forever be resigned to the confines of the lab and will never make it into  
1008 real-world applications. This will require a concentrated effort to export affect detection  
1009 out of the lab and into the wild, where one must contend with the dynamic nature and  
1010 unpredictability of the real world. It is our hope that this will be reflected in the next  
1011 review of multimodal affect detectors.

1012 **REFERENCES**

- 1013 S. Afzal and P. Robinson. 2011. Natural affect data: Collection and annotation. In *New Perspectives on Affect*  
1014 *and Learning Technologies*, R. Calvo and S. D’Mello (Eds.) Springer, New York, NY, 44–70.

## A Review and Meta-Analysis of Multimodal Affect Detection Systems

43:31

- J. Bailenson, E. Pontikakis, I. Mauss, J. Gross, M. Jabon, C. Hutcherson, C. Nass, and O. John. 2008. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *Int. J. Hum. Comput. Stud.* 66, 303–317. 1015–1017
- T. Baltrušaitis, N. Banda, and P. Robinson. 2013. Dimensional affect recognition using continuous conditional random fields. In *Proceedings of the International Conference on Multimedia and Expo (Workshop on Affective Analysis in Multimedia)*. 1018–1020
- N. Banda and P. Robinson. 2011. Noise analysis in audio-visual emotion recognition. In *Proceedings of the 11th International Conference on Multimodal Interaction (ICMI)*. 1021–1022
- L. Barrett. 2006. Are emotions natural kinds? *Perspect. Psychol. Sci.* 1, 28–58. 1023
- L. Barrett, B. Mesquita, K. Ochsner, and J. Gross. 2007. The experience of emotion. *Ann. Rev. Psychol.* 58, 373–403. 1024–1025
- M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. 2009. *Introduction to Meta-Analysis*. John Wiley & Sons, Inc., Hoboken, NJ. 1026–1027
- S. Brave and C. Nass. 2002. Emotion in human-computer interaction. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. Jacko and A. Sears (Eds.). Erlbaum Associates, Inc., Hillsdale, NJ, 81–96. 1028–1030
- C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI'04)*, R. Sharma, T. M. P. Darrell Harper, G. Lazzari and M. Turk (Eds.). ACM, State College, PA, 205–211. 1031–1034
- R. Calvo, S. K. D'Mello, J. Gratch, and A. Kappas. 2014. *The Oxford Handbook of Affective Computing*. Oxford University Press, New York, NY. 1035–1036
- R. A. Calvo and S. K. D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affect. Comput.* 1 (2007), 18–37. 1037–1038
- G. Caridakis, L. Malatesta, L. Kessous, N. Amir, A. Paouzaoui, and K. Karpouzis. 2006. Modeling naturalistic affective states via facial and vocal expression recognition. In *International Conference on Multimodal Interfaces*. ACM, New York, NY, 146–154. 1039–1041
- G. Castellano, L. Kessous, and G. Caridakis. 2008. Emotion recognition through multiple modalities: Face, body gesture, speech. In *Affect and Emotion in Human-Computer Interaction*, C. Peter and R. Beale (Eds.). Lecture Notes in Computer Science, Vol. 4868. Springer, Berlin, 92–103. 1042–1044
- G. Castellano, A. Pereira, I. Leite, A. Paiva, and P. McOwan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features. In *Proceedings of the 2009 International Conference on Multimodal interfaces*. ACM, New York, NY, 119–126. 1045–1047
- G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun. 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Trans. Syst., Man Cybern. Part A Syst. Humans* 41, 1052–1063. 1048–1049
- C.-Y. Chen, Y.-K. Huang, and P. Cook. 2005. Visual/Acoustic emotion recognition. In *Proceedings of the IEEE International Conference on Multimedia and Expo*. IEEE, Washington, DC, 1468–1471. 1050–1051
- G. Chetty and M. Wagner. 2008. A multilevel fusion approach for audiovisual emotion recognition. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 115–120. 1052–1053
- Z.-J. Chuang and C.-H. Wu. 2004. Multi-modal emotion recognition from speech and text. *Int. J. Comput. Ling. Chin. Lang. Process.* 9, 1–18. 1054–1055
- J. A. Coan. 2010. Emergent ghosts of the emotion machine. *Emotion Rev.* 2, 274–285. 1056
- J. Cohen. 1992. A power primer. *Psychol. Bull.* 112, 155–159. 1057
- C. Conati and H. Maclaren. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Model. User-Adapt. Interact.* 19, 267–303. 1058–1059
- C. Conati, S. Marsella, and A. Paiva. 2005. Affective interactions: The computer in the affective loop. In *Proceedings of the 10th International Conference on Intelligent User Interfaces*, J. Riedl and A. Jameson (Eds.). ACM, New York, NY, 7. 1060–1062
- R. Cowie, E. Douglas-Cowie, and C. Cox. 2005. Beyond emotion archetypes: Databases for emotion modelling using neural networks. *Neur. Netw.* 18, 371–388. 1063–1064
- R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Sig. Process. Mag.* 18, 32–80. 1065–1066
- D. Cueva, R. Gonçalves, F. Cozman, and M. Pereira-Barretto. 2011. Crawling to improve multimodal emotion detection. In *Proceedings of the 10th Mexican International Conference on Artificial Intelligence (MICAI'11)*. Springer-Verlag, Puebla, Mexico, 343–350. 1067–1068
- S. D'Mello. 2013. A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *J. Educ. Psychology Psychol.* 105, 1082–1099. 1069–1070–1071

- 1072 S. D'Mello and A. Graesser. 2007. Mind and body: Dialogue and posture for affect detection in learning  
1073 environments. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*,  
1074 R. Lukin et al. (Eds.). IOS Press, Amsterdam, 161–168.
- 1075 S. D'Mello and A. Graesser. 2010. Multimodal semi-automated affect detection from conversational cues,  
1076 gross body language, and facial features. *User Model. User-Adap. Interact.* 20, 147–187.
- 1077 S. D'Mello and A. Graesser. 2012. AutoTutor and affective autotutor: Learning by talking with cognitively  
1078 and emotionally intelligent computers that talk back. *ACM Trans. Interact. Intell. Syst.* 2, 23:22–23:39.
- 1079 S. D'Mello and J. Kory. 2012. Consistent but modest: Comparing multimodal and unimodal affect detection  
1080 accuracies from 30 studies. In *Proceedings of the 14th ACM International Conference on Multimodal  
1081 Interaction*, L.-P. Morency, D. Bohus, H. Aghajan, A. Nijholt, J. Cassell and J. Epps (Eds.). ACM New  
1082 York, NY, 31–38.
- 1083 S. K. D'Mello and A. C. Graesser. 2014. Confusion. In *International Handbook of Emotions in Education*,  
1084 R. Pekrun and L. Linnenbrink-Garcia (Eds.). Routledge, New York, NY, 289–310.
- 1085 S. D'Mello and A. Graesser. 2011. The half-life of cognitive-affective states during complex learning. *Cognition  
1086 Emotion* 25, 1299–1308.
- 1087 D. Datcu and L. Rothkrantz. 2011. Emotion recognition using bimodal data fusion. In *Proceedings of the  
1088 12th International Conference on Computer Systems and Technologies*. ACM, New York, NY, 122–128.
- 1089 S. Dobrišek, R. Gajšek, F. Mihelič, N. Pavešič, and V. Štruc. 2013. Towards efficient multi-modal emotion  
1090 recognition. *Int. J. Adv. Robotic Syst.* 10, 1–10.
- 1091 E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. C. Martin, L. Devillers, S. Abrilian,  
1092 and A. Batliner. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic  
1093 and induced emotional data. In *Proceedings of the 2nd International Conference on Affective Computing  
1094 and Intelligent Interaction*. Springer, Berlin, 488–500.
- 1095 S. Duval and R. Tweedie. 2000. Trim and fill: A simple funnel-plot-based method of testing and adjusting  
1096 for publication bias in meta-analysis. *Biometrics* 56, 455–463.
- 1097 M. Dy, I. Espinosa, P. Go, C. Mendez, and J. Cu. 2010. Multimodal emotion recognition using a sponta-  
1098 neous Filipino emotion database. In *Proceedings of the 3rd International Conference on Human-Centric  
1099 Computing*. IEEE, Washington, DC, 1–5.
- 1100 P. Ekman. 1992. An argument for basic emotions. *Cognition Emotion* 6, 169–200.
- 1101 H. Elfенbein and N. Ambady. 2002. On the universality and cultural specificity of emotion recognition: A  
1102 meta-analysis. *Psychol. Bull.* 128, 203–235.
- 1103 S. Emerich, E. Lupu, and A. Apatean. 2009. Emotions recognition by speech and facial expressions analysis.  
1104 In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009)*. Glasgow, Scotland.
- 1105 F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie. 2010. On-line emotion  
1106 recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *J. Multimodal  
1107 User Int.* 3, 7–19.
- 1108 F. Eyben, M. Wollmer, M. F. Valstar, H. Gunes, B. Schuller, and M. Pantic. 2011. String-based audiovisual  
1109 fusion of behavioural events for the assessment of dimensional affect. In *Ninth IEEE International  
1110 Conference on Automatic Face and Gesture Recognition (FG 2011)*. IEEE, Santa Barbara, CA, 322–329.
- 1111 J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth. 2007. The world of emotions is not two-dimensional.  
1112 *Psychol. Sci.* 18, 12 (Dec. 2007) 1050–1057.
- 1113 K. Forbes-Riley and D. Litman. 2004. Predicting emotion in spoken dialogue from multiple knowledge  
1114 sources. In *Proceedings of the 4th Meeting of the North American Chapter of the Association for Compu-  
1115 tational Linguistics: Human Language Technologies*, 201–208.
- 1116 K. Forbes-Riley and D. J. Litman. 2011. Benefits and challenges of real-time uncertainty detection and  
1117 adaptation in a spoken dialogue computer tutor. *Speech Commun.* 53, 1115–1136.
- 1118 R. Gajšek, V. Štruc, and F. Mihelič. 2010. Multi-modal emotion recognition using canonical correlations and  
1119 acoustic features. In *Proceedings of the 20th International Conference on Pattern Recognition*. IEEE,  
1120 Washington, DC, 4133–4136.
- 1121 M. Glodek, S. Reuter, M. Schels, K. Dietmayer, and F. Schwenker. 2013. Kalman filter based classifier fusion  
1122 for affective state recognition. In *Proceedings of the 11th International Workshop on Multiple Classifier  
1123 Systems*, Z.-H. Zhou, F. Roli, and J. Kittler (Eds.). Springer, Berlin, 85–94.
- 1124 M. Glodek, S. Tschechne, G. Layher, M. Schels, T. Brosch, S. Scherer, M. Kächele, M. Schmidt, H. Neumann,  
1125 and G. Palm. 2011. Multiple classifier systems for the classification of audio-visual emotional states. In  
1126 *4th International Conference on Affective Computing and Intelligent Interaction (ACII'11)*, S. D'Mello, A.  
1127 Graesser, B. Schuller, and J. Martin (Eds.). Springer, Memphis, TN, 359–368.
- 1128 S. Gong, C. Shan, and T. Xiang. 2007. Visual inference of human emotion and behaviour. In *Proceedings of  
1129 the 9th International Conference on Multimodal Interfaces*. ACM, New York, NY, 22–29.



## A Review and Meta-Analysis of Multimodal Affect Detection Systems 43:33

- A. Graesser, B. McDaniel, P. Chipman, A. Witherspoon, S. D’Mello, and B. Gholson. 2006. Detection of emotions during learning with AutoTutor. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, R. Sun and N. Miyake (Eds.). Cognitive Science Society, Austin, TX, 285–290. 1130–1132
- H. Gunes and M. Piccardi. 2005. Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction (ACII’05)*, J. Tao and R. Picard (Eds.). Springer-Verlag, 102–111. 1133–1136
- H. Gunes and M. Piccardi. 2009. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Trans. Syst., Man, Cybern. Part B Cybern.* 39, 64–84. 1137–1138
- M. Han, J. Hsu, K.-T. Song, and F.-Y. Chang. 2007. A new information fusion method for SVM-based robotic audio-visual emotion recognition. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. IEEE, Washington, DC, 2656–2661. 1139–1141
- S. Haq and P. Jackson. 2009. Speaker-dependent audio-visual emotion recognition. In *Proceedings of International Conference on Auditory-Visual Speech Processing*, 53–58. 1142–1143
- S. Haq, P. Jackson, and J. Edge. 2008. Audio-visual feature selection and reduction for emotion classification. In *Proceedings of the International Conference on Auditory-Visual Speech Processing*, 185–190. 1144–1145
- S. Hoch, F. Althoff, G. McGlaun, and G. Rigoll. 2005. Bimodal fusion of emotional data in an automotive environment. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Washington, DC, 1085–1088. 1146–1148
- S. Hommel, A. Rabie, and U. Handmann. 2013. Attention and emotion based adaptation of dialog systems. In *Intelligent Systems: Models and Applications*, E. Pap (Ed.). Springer-Verlag, Berlin, 215–235. 1149–1150
- M. Hoque and R. W. Picard. 2011. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG’11)*. IEEE, Washington, DC, 354–359. 1151–1153
- M. Hussain, H. Monkaresi, and R. Calvo. 2012. Combining classifiers in multimodal affect detection. In *Proceedings of the Australasian Data Mining Conference*. 1154–1155
- C. Izard. 2010. The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Rev.* 2, 363–370. 1156–1157
- C. E. Izard. 2007. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspect. Psychol. Sci.* 2, 260–280. 1158–1159
- A. Jaimes and N. Sebe. 2007. Multimodal human-computer interaction: A survey. *Comput. Vision Image Understanding* 108, 116–134. 1160–1161
- L. Jeni, J. Cohn, and F. De La Torre. 2013. Facing imbalanced data—Recommendations for the use of performance metrics. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII’13)*, A. Nijholt, S. K. D’Mello, and M. Pantic (Eds.). IEEE, Washington, DC, 245–251. 1162–1165
- D. Jiang, Y. Cui, X. Zhang, P. Fan, I. Ganzalez, and H. Sahli. 2011. Audio visual emotion recognition based on triple-stream dynamic bayesian network models. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, S. D’Mello, A. Graesser, B. Schuller, and J. Martin (Eds.). Springer-Verlag, 609–618. 1166–1169
- J.-T. Joo, S.-W. Seo, K.-E. Ko, and K.-B. Sim. 2007. Emotion recognition method based on multimodal sensor fusion algorithm. *한국사능시스템학회 국제 학술대회 발표논문집*, 200–204. 1170–1171
- C. Kaernbach. 2011. On dimensions in emotion psychology. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition and Workshops*. IEEE, Washington, DC, 792–796. 1172–1173
- I. Kanluan, M. Grimm, and K. Kroschel. 2008. Audio-visual emotion recognition using an emotion space concept. In *Proceedings of the 16th European Signal Processing Conference*. 1174–1175
- A. Kapoor, B. Burleson, and R. Picard. 2007. Automatic prediction of frustration. *Int. J. Hum. Comput. Stud.* 65, 724–736. 1176–1177
- A. Kapoor and R. Picard. 2005. Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM International Conference on Multimedia*. ACM, New York, NY, 677–682. 1178–1179
- K. Karpouzis, G. Caridakis, L. Kessous, N. Amir, A. Raouzaoui, L. Malatesta, and S. Kollias. 2007. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In *Artificial Intelligence for Human Computing*, T. Huang (Ed.). Springer-Verlag, Berlin, 91–112. 1180–1182
- L. Kessous, G. Castellano, and G. Caridakis. 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *J. Multimodal User Int.* 3, 33–48. 1183–1184
- Z. Khalali and M. Moradi. 2009. Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of EEG. In *Proceedings of International Joint Conference on Neural Networks*. IEEE, Los Alamitos, CA, 1571–1575. 1185–1186

- 1188 J. Kim. 2007. Bimodal emotion recognition using speech and physiological changes. In *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel (Eds.). I-Tech, 265–280.
- 1189
- 1190 J. Kim, E. André, M. Rehm, T. Vogt, and J. Wagner. 2005. Integrating information from speech and physiological signals to achieve emotional sensitivity. In *Proceedings of 9th European Conference on Speech Communication and Technology*. 809–812.
- 1191
- 1192
- 1193 J. Kim and F. Lingenfelter. 2010. Ensemble approaches to parametric decision fusion for bimodal emotion recognition. In *Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*. BIOSTEC, 460–463.
- 1194
- 1195
- 1196 S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. Deap: A database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31.
- 1197
- 1198
- 1199 J. Kory and S. K. D'Mello. 2014. 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.
- 1200
- 1201
- 1202 G. Krell, M. Glodek, A. Panning, I. Siegert, B. Michaelis, A. Wendemuth, and F. Schwenker. 2013. Fusion of fragmentary classifier decisions for affective state recognition. In *Proceedings of 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, 116–130.
- 1203
- 1204
- 1205 M. D. Lewis. 2005. Bridging emotion theory and neurobiology through dynamic systems modeling. *Behav. Brain Sci.* 28, 169–245.
- 1206
- 1207
- 1208 J. Lin, C. Wu, and W. Wei. 2012. Error weighted semi-coupled hidden markov model for audio-visual emotion recognition. *IEEE Trans. Multimedia* 14, 142–156.
- 1209
- 1210 F. Lingenfelter, J. Wagner, and E. André. 2011. 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.
- 1211
- 1212
- 1213 M. W. Lipsey and D. B. Wilson. 2001. *Practical meta-analysis*. Sage Publications, Inc, Thousand Oaks, CA.
- 1214
- 1215 D. Litman and K. Forbes-Riley. 2004. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Barcelona, Spain, 352–359.
- 1216
- 1217 D. Litman and K. Forbes-Riley. 2006a. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Commun.* 48, 559–590.
- 1218
- 1219 D. J. Litman and K. Forbes-Riley. 2006b. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Commun.* 48, 559–590.
- 1220
- 1221
- 1222 K. Lu and Y. Jia. 2012. Audio-visual emotion recognition with boosted coupled HMM. In *Proceedings of the 21st International Conference on Pattern Recognition*. IEEE, Washington, DC, 1148–1151.
- 1223
- 1224 M. Mansoorizadeh and N. Charkari. 2010. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools Appl.* 49, 277–297.
- 1225
- 1226 D. McDuff, R. Kaliouby, and R. W. Picard. 2012. Crowdsourcing facial responses to online videos. *IEEE Trans. Affective Comput.* 3, 456–468.
- 1227
- 1228 G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. 2012. The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Trans. Affective Comput.* 3, 5–17.
- 1229
- 1230
- 1231 A. Metallinou, S. Lee, and S. Narayanan. 2008. Audio-visual emotion recognition using Gaussian mixture models for face and voice. In *Proceedings of the 10th IEEE International Symposium on Multimedia*. IEEE, Washington, DC, 250–257.
- 1232
- 1233
- 1234 A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan. 2012. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affective Comput.* 3, 184–198.
- 1235
- 1236 H. Monkaresi, M. S. Hussain, and R. Calvo. 2012. 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.
- 1237
- 1238
- 1239 M. Nicolaou, H. Gunes, and M. Pantic. 2011. Continuous prediction of spontaneous affect from multiple cues and modalities in valence and arousal space. *IEEE Trans. Affective Comput.* 2, 92–105.
- 1240
- 1241 J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan, and C. Heffernan. 2014. Population validity for educational data mining: A case study in affect detection. *Brit. J. Educ. Psychol.* 45, 487–501.
- 1242
- 1243 A. Ortony, G. Clore, and A. Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press, New York.
- 1244

## A Review and Meta-Analysis of Multimodal Affect Detection Systems

43:35

- P. Pal, A. Iyer, and R. Yantorno. 2006. Emotion detection from infant facial expressions and cries. In *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, Washington, DC, 721–724. 1245  
1246  
1247
- M. Paleari, R. Benmokhtar, and B. Huet. 2009. Evidence theory-based multimodal emotion recognition. 1248  
In *Proceedings of the 15th International Multimedia Modeling Conference (MMM'09)*. Springer-Verlag, 435–446. 1249  
1250
- B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retrieval* 2, 1–135. 1251
- M. Pantic and L. Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* 91, 1370–1390. 1252  
1253
- J. Park, G. Jang, and Y. Seo. 2012. Music-aided affective interaction between human and service robot. *EURASIP J. Audio, Speech, Music Process.* 2012, 1, 1–13. 1254  
1255
- R. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, Mass. 1256
- R. Picard. 2010. Affective Computing: From Laughter to IEEE. *IEEE Trans. Affective Comput.* 1, 11–17. 1257
- R. Plutchik. 2001. The nature of emotions. *American Scientist* 89, 344–350. 1258
- A. Rabie, B. Wrede, T. Vogt, and M. Hanheide. 2009. Evaluation and discussion of multi-modal emotion recognition. In *Proceedings of the Second International Conference on Computer and Electrical Engineering (ICCEE'09)*. IEEE Computer Society, 598–602. 1259  
1260  
1261
- M. Rashid, S. Abu-Bakar, and M. Mokji. 2012. Human emotion recognition from videos using spatio-temporal and audio features. *Visual Comput.* 29, 12, 1269–1275. 1262  
1263
- G. Rigoll, R. Muller, and B. Schuller. 2005. Speech emotion recognition exploiting acoustic and linguistic information sources. In *Proceedings of the 10th International Conference Speech and Computer*. 61–67. 1264  
1265
- V. Rosas, R. Mihalcea, and L. Morency. 2013. Multimodal sentiment analysis of spanish online videos. *IEEE Intell. Syst.* 1266  
1267
- E. Rosenberg. 1998. Levels of analysis and the organization of affect. *Rev. Gen. Psychol.* 2, 247–270. 1268
- E. Rosenberg and P. Ekman. 1994. Coherence between expressive and experiential systems in emotion. *Cognition Emotion* 8, 201–229. 1269  
1270
- V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad. 2012. Ensemble of SVM trees for multimodal emotion recognition. In *Proceedings of the Signal and Information Processing Association Annual Summit and Conference*. IEEE, Washington, DC, 1–4. 1271  
1272  
1273
- J. Russell. 1994. Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychol. Bull.* 115, 102–141. 1274  
1275
- J. Russell. 2003. Core affect and the psychological construction of emotion. *Psychol. Rev.* 110, 145–172. 1276
- J. A. Russell, J. A. Bachorowski, and J. M. Fernandez-Dols. 2003. Facial and vocal expressions of emotion. *Ann. Rev. Psychol.* 54, 329–349. 1277  
1278
- A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma. 2012. 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 Press, New York, NY, 485–492. 1279  
1280  
1281
- B. Schuller. 2011. Recognizing affect from linguistic information in 3D continuous space. *IEEE Trans. Affective Comput.* 2, 192–205. 1282  
1283
- B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu, and G. Rigoll. 2007. Audiovisual recognition of spontaneous interest within conversations. In *Proceedings of the 9th International Conference on Multimodal Interfaces*. ACM, New York, NY, 30–37. 1284  
1285  
1286
- N. Sebe, I. Cohen, T. Gevers, and T. Huang. 2006. Emotion recognition based on joint visual and audio cues. In *Proceedings of the 18th International Conference on Pattern Recognition*. IEEE, Washington, DC, 1136–1139. 1287  
1288  
1289
- D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson. 2008. Patterns, prototypes, performance: Classifying emotional user states. In *Proceedings of the 9th Annual Conference of the International Speech Communication Association*, 601–604. 1290  
1291  
1292
- C. Shan, S. Gong, and P. McOwan. 2007. Beyond facial expressions: Learning human emotion from body gestures. In *Proceedings of the British Machine Vision Conference*, 1–10. 1293  
1294
- M. Soleymani, M. Pantic, and T. Pun. 2012. Multi-modal emotion recognition in response to videos. *IEEE Trans. Affective Comput.* 3, 211–223. 1295  
1296
- A. Sutcliffe. 2008. Multimedia user interface design. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, A. Sears and J. Jacko (Eds.). Taylor & Francis, New York, NY, 245–261. 1297  
1298  
1299
- S. S. Tomkins. 1962. *Affect Imagery Consciousness: Volume I, The Positive Affects*. Tavistock, London. 1300

- 1301 B. Tu and F. Yu. 2012. Bimodal emotion recognition based on speech signals and facial expression. In  
 1302 *Proceedings of the 6th International Conference on Intelligent Systems and Knowledge*. Springer, Berlin,  
 1303 691–696.
- 1304 J. Tukey and D. McLaughlin. 1963. Less vulnerable confidence and significance procedures for location based  
 1305 on a single sample: Trimming/Winsorization 1. *Sankhyā: The Indian Journal of Statistics* 25, 331–352.
- 1306 M. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. 2012. Meta-analysis of the first facial expression  
 1307 recognition challenge. *IEEE Trans. Syst., Man, Cybern. Part B Cybern.* 42, 966–979.
- 1308 M. van der Zwaag, J. Janssen, and J. Westerink. 2013. Directing physiology and mood through music:  
 1309 Validation of an affective music player. *IEEE Trans. Affective Comput.* 4, 57–68.
- 1310 H. Vu, Y. Yamazaki, F. Dong, and K. Hirota. 2011. Emotion recognition based on human gesture and  
 1311 speech information using RT middleware. In *IEEE International Conference on Fuzzy Systems*. IEEE,  
 1312 Washington, DC, 787–791.
- 1313 J. Wagner, E. Andre, F. Lingenfeller, J. Kim, and T. Vogt. 2011. Exploring fusion methods for multimodal  
 1314 emotion recognition with missing data. *IEEE Trans. Affective Comput.* 2, 206–218.
- 1315 S. Walter, S. Scherer, M. Schels, M. Glodek, D. Hrabal, M. Schmidt, R. Böck, K. Limbrecht, H. Traue, and  
 1316 F. Schwenker. 2011. Multimodal emotion classification in naturalistic user behavior. In *Proceedings of*  
 1317 *the International Conference on Human-Computer Interaction*, J. Jacko (Ed.). Springer, Berlin, 603–611.
- 1318 S. Wang, Y. Zhu, G. Wu, and Q. Ji. 2013. Hybrid video emotional tagging using users' EEG and video content.  
 1319 *Multimed. Tools Appl.* 1–27.
- 1320 Y. Wang and L. Guan. 2005. Recognizing human emotion from audiovisual information. In *Proceedings of*  
 1321 *the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Washington, DC,  
 1322 1125–1128.
- 1323 Y. Wang and L. Guan. 2008. Recognizing human emotional state from audiovisual signals. *IEEE Trans.*  
 1324 *Multimedia* 10, 936–946.
- 1325 M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig. 2008. Low-level fusion of audio and video feature  
 1326 for multi-modal emotion recognition. In *Proceedings of the 3rd International Conference on Computer*  
 1327 *Vision Theory and Applications*, 145–151.
- 1328 M. Wöllmer, M. Kaiser, F. Eyben, and B. Schuller. 2013a. LSTM modeling of continuous emotions in an  
 1329 audiovisual affect recognition framework. *Image Vision Comput.* 31, 2 (Feb. 2013), 153–163.
- 1330 M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan. 2010. Context-sensitive multi-  
 1331 modal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In  
 1332 *Proceedings of the 11th Annual Conference of the International Speech Communication Association (IN-*  
 1333 *TERSPEECH'10)*. 2362–2365.
- 1334 M. Wöllmer, F. Wening, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L. Morency. 2013b. YouTube movie  
 1335 reviews: In, cross, and open-domain sentiment analysis in an audiovisual context. *IEEE Intell. Syst.*
- 1336 C. Wu and W. Liang. 2011. Emotion recognition of affective speech based on multiple classifiers using  
 1337 acoustic-prosodic information and semantic labels. *IEEE Trans. Affective Comput.* 2, 10–21.
- 1338 Z. Zeng, Y. Hu, Y. Fu, T. Huang, G. Roisman, and Z. Wen. 2006. Audio-visual emotion recognition in adult  
 1339 attachment interview. In *Proceedings of the 8th International Conference on Multimodal Interfaces*.  
 1340 ACM, Washington, DC, 139–145.
- 1341 Z. Zeng, M. Pantic, G. Roisman, and T. Huang. 2009. A survey of affect recognition methods: Audio, visual,  
 1342 and spontaneous expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 39–58.
- 1343 Z. Zeng, J. Tu, M. Liu, and T. Huang. 2005. Multi-stream confidence analysis for audio-visual affect recogni-  
 1344 tion. In *Proceedings of the 1st International Conference on Affective Computing and Intelligent Interac-*  
 1345 *tion*, J. Tao., T. Tan. and R. Picard. (Eds.). Springer, Berlin, 964–971.
- 1346 Z. Zeng, J. Tu, M. Liu, T. Huang, B. Pianfetti, D. Roth, and S. Levinson. 2007. Audio-visual affect recognition.  
 1347 *IEEE Trans. Multimedia* 9, 424–428.

**Q11** Received xxxx; revised xxxx; accepted xxxx.

## QUERIES

- Q1:** AU: Please check symbol.
- Q2:** AU: Please check symbol.
- Q3:** AU: Please check change to em dash.
- Q4:** AU: Please check journal abbreviation.
- Q5:** AU: Please check journal abbreviation.
- Q6:** AU: Please check all information in Fontaine et al. 2007.
- Q7:** AU: Please check CJK characters for accuracy.
- Q8:** AU: Please check vol., issue and page nos.
- Q9:** AU: Please update vol. and pg. range if possible.
- Q10:** AU: Please provide volume and page range if possible.
- Q11:** AU: Please provide history dates.