Consistent but Modest: A Meta-Analysis on Unimodal and Multimodal Affect Detection Accuracies from 30 Studies

Sidney D'Mello
University of Notre Dame
Notre Dame, IN 46556, USA
(011)-574-631-1822
sdmello@nd.edu

Jacqueline Kory
University of Notre Dame
Notre Dame, IN 46556, USA
(011)-574-631-1822
jkory@nd.edu

ABSTRACT
The recent influx of multimodal affect classifiers raises the important question of whether these classifiers yield accuracy rates that exceed their unimodal counterparts. This question was addressed by performing a meta-analysis on 30 published studies that reported both multimodal and unimodal affect detection accuracies. The results indicated that multimodal accuracies were consistently better than unimodal accuracies and yielded an average 8.12% improvement over the best unimodal classifiers. However, performance improvements were three times lower when classifiers were trained on natural or seminatural data (4.39% improvement) compared to acted data (12.1% improvement). Importantly, performance of the best unimodal classifier explained an impressive 80.6% (cross-validated) of the variance in multimodal accuracy. The results also indicated that multimodal accuracies were substantially higher than accuracies of the second-best unimodal classifiers (an average improvement of 29.4%) irrespective of the naturalness of the training data. Theoretical and applied implications of the findings are discussed.

Categories and Subject Descriptors
H.5.m **[Information Interfaces and Presentation]:** Miscellaneous.

Keywords
Affect detection, emotion detection, affective computing, multimodal affect detection, meta-analysis, review

1. INTRODUCTION
Affect detection has been, and continues to be, on the forefront of Affective Computing (AC) research. The emphasis on affect detection is reasonable because a system can never respond to users' emotions if it cannot detect their emotions. Consequently, the last 15 years have witnessed 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 [1-3] for reviews). Although some early affect detection systems focused primarily on individual modalities and on emotional expressions portrayed by actors, several of the contemporary systems emphasize multimodal detection of naturalistic affective expressions. This suggests that the field is moving in the right direction.

ICMI '12, October 22–26, 2012, Santa Monica, California, USA.

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. The field is still confronted with the persistent problems of: (a) intrusive, expensive, and noisy sensors that are largely unscalable, (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 and realistic training data for machine learning models, (d) challenges of incorporating top-down models of context and appraisals with bottom-up bodily- and physiological-based sensing, (e) lack of clarity of the affective phenomenon being modeled (e.g., moods vs. emotions, categorical vs. dimensional representations), (f) issues pertaining to generalizability across contexts, time, individual differences, and cultural differences, and (g) many others.

As AC researchers are well aware, this laundry 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. Numerous innovative solutions are being developed to address several of these issues. One strategy that is gaining momentum in the literature is to alleviate the noisy signal problem (item b from the list above) by increasing the amount of available data. More specifically, several researchers are focusing on developing multimodal affect detectors with the assumption that incorporating multimodal signals will yield classification accuracies that are superior to unimodal signals. Although this assumption has obvious face validity, the results of these endeavors are somewhat mixed. When compared to the accuracies obtained by the best unimodal classifiers, some studies have reported impressive multimodal improvements (e.g., [4-8]), others have reported negligible or null improvements (e.g., [9-11]), and some have even reported negative improvements (e.g., [12-14]).

The considerable inter-study variance in the results of multimodal affect classifiers makes it difficult to appropriately gauge what advantages (if any) multimodal classification yields over unimodal classification. More importantly, is it possible to identify situations where multimodal classifiers are expected to yield impressive improvements and differentiate these from situations that result in null or negative improvements? The present paper makes an initial attempt to address these questions by analyzing multimodal and unimodal classification accuracies reported in 30 published affect detection studies.

The analyses focused on quantifying the *added value* afforded by multimodal affect classification above and beyond unimodal classification (called MM effect size). We were also interested in identifying factors that reliably predicted MM effects. Our focus is on quantifying performance rather than analyzing the different approaches used to integrate multimodal information (see [1-3]

for reviews on different multimodal systems). As such, this paper should be considered to be more of a meta-analysis of the multimodal affect detection literature rather than a review or survey paper. Hence, the descriptions of the studies themselves are quite brief (Section 2), but we emphasize the data (Section 3) and various analyses on the data (Section 4).

2. SUMMARY OF STUDIES ANALYZED

Studies were selected by formally searching relevant journals (e.g., *IEEE Transactions of Affective Computing, IEEE Transactions on Multimedia*) and conference proceedings (e.g., *Affective Computing and Intelligent Interaction, Multimodal Interaction, Face and Gesture*). Informal searches for variants of the terms *"multimodal" and "affect or emotion" and "detection or classification"* were also performed via Google Scholar. Any peer-reviewed publication that reported both unimodal and multimodal 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. More recent studies were given preference over older studies. Selection bias was avoided by including the first 30 studies that satisfied these basic criteria. A summary of the studies along several dimensions is presented below.

2.1 Data type

This pertains to whether the data used to train and validate the unimodal and multimodal classifiers consisted of emotional expressions that were: (a) obtained by asking *actors* to portray various emotions (e.g., [8, 11, 12, 15-19]), (b) collected via experimental methods that *induced* specific emotions (e.g., [9, 14, 20]), (c) *naturalistic* displays of emotion (e.g., [21-24]), or (d) some blend of (b) and (c) (e.g., [13, 25-27]), henceforth referred to as *seminatural.* Approximately half of the multimodal classifiers (53.3%) were trained and validated on acted data, 20% used semi-natural data, 16.6% used natural data, and 10% induced specific emotions.

While the criteria for a dataset to be categorized as acted, natural, or induced is quite clear, the seminatural category requires some clarification. This designation was mainly applied to data sets where specific emotions were not directly induced (e.g., showing participants emotion-induction films [20]), but the interaction itself was intentionally emotionally charged, thereby increasing the likelihood that participants would respond emotionally. For example, the SEMAINE database [28] was constructed by asking participants to engage in a conversation with an avatar that had one of four affective dispositions (or personalities): angry, happy, gloomy, or pragmatic. Studies that utilized this database (e.g., [27, 29]) were categorized as seminatural because it is likely that the personality of the avatar influenced the emotions of the participant. In particular, there is considerable evidence that perceiving an emotion increases the likelihood of experiencing that emotion (i.e., emotion contagion) because of shared neural substrates that underlie perception and expression [30].

2.2 Classification task

This refers to whether the affect detector performed a categorical versus a dimensional classification. *Categorical* classification consists of discriminating between one of *k* affect labels (e.g., anger, sadness, happiness, and neutral [25]). *Dimensional* classification involved predicting activity on a particular affective dimension (e.g., valence, activation, dominance [24]). There were a number of studies that used a dimensional model to annotate affect, but proceeded with a categorical classification by dividing the dimensional affective space. These have been categorized as *mixed*. For example, Wöllmer [6] used 5 point scales to annotate valence and activation, but then performed a categorical classification by performing a tripartite split on each dimension (i.e., dividing the scale into low, medium, and high sections). Similarly, continuous activation-valence values were discretized by clustering prior to classification in [27]. A vast majority (66.7%) of the studies performed a categorical classification, while mixed (23.3%) and dimensional (10%) classifications were comparatively rare.

2.3 Affective states classified

Most of the studies focused on all or some subset of the "basic" emotions (e.g., [7, 8, 12, 25, 31]). Several studies focused on predicting valence and arousal, usually independently (e.g., [6, 10, 13, 20, 32]), but occasionally jointly in a valence-arousal space (e.g., [9, 16, 27]). Studies that focused on non-basic emotions (e.g., [22]) or on some of the less studied dimensions of expectancy and power (e.g., [13, 26]) were less frequent.

2.4 Modalities

Most of the studies analyzed facial expressions (77%) and acoustic-prosodic cues (77%). Almost a third (30%) of the studies tracked some form of body movements, postures, and gestures. Text, EEG, biosignals (i.e., ECG, EMG, GSR, etc.), eye gaze, event, and context were comparatively rare. Accordingly, audio-visual features constituted the most common multimodal systems (43.3%) followed by a trimodal face + voice + posture and gesture systems (16.7%). Overall, 73.3% of the multimodal systems were bimodal, while 26.7% were trimodal.

2.5 Multimodal Fusion Techniques

Most studies considered different multiple fusion techniques, so it is difficult to accurately estimate if a particular method or technique was more commonly used than the others. In general, many studies compared naïve feature-based fusion with decision-level fusion (e.g., [4, 12, 19, 20]), model-level fusion [7, 10, 29], and more sophisticated fusion strategies. These include mixtures of Gaussian process classification [23], string-based approaches [26], hybrid approaches that combine feature-level and decision-level fusion [9] or decision-level and model-level fusion [7], bidirectional long short-term memory neural networks [10, 29], emotion-adapted decision-level fusion [16], and meta-decision trees [31].

3. THE DATA

Table 1 provides both unimodal and multimodal classification performance scores. The 27 studies that performed a categorical classification used classification accuracy (i.e., the proportion of correctly classified instances) as the evaluation metric. In rare cases where both classification accuracy and the F1-measure was reported (e.g., [6, 10]), classification accuracy was taken to be the metric in order to increase consistency among studies. The dimensional studies typically reported a correlation coefficient, and this was taken as the performance metric.

The rightmost column (MM Effect) is the metric used to quantify multimodal performance improvement as a function of unimodal performance. If a_1 and a_2 are accuracies associated with two unimodal classifiers, and a_{12} is the multimodal accuracy, then the multimodal effect is $100 * \frac{a_{12} - max(a_1, a_2)}{max(a_1, a_2)}$. This is simply the percent improvement over the best unimodal classifier. This metric affords a unified analysis of studies that use classification accuracy and the correlation coefficient as the evaluation metric.

It is important to note two points about the data presented in Table 1. First, accuracy scores associated with the best performing classifier were used in situations where multiple classifiers were considered for the *same* classification task. For example, [20] reported both feature-level and decision-level multimodal classification accuracy rates. Decision-level fusion yielded higher accuracy rates, so only decision level fusion results were used for in the subsequent analyses.

Second, there were more data points (N = 47) than studies (N = 30) because some studies performed *multiple* classification tasks. For example, [22] developed one classifier to predict four affective states and another to predict an overlapping but different set of five affective states. Other than this exception, in general, one data point was obtained for the studies that performed a categorical classification. It was the dimensional studies that contributed multiple data points because the number of classification models increases linearly with the number of dimensions considered. For example, the study by Eyben and colleagues [26] contributed five data points because their models independently predicted five affect dimensions (i.e., activation, expectancy, intensity, power, and valence).

There was the concern that studies that contributed multiple data points would bias the MM effect distributions and the summary statistics (i.e., means and standard deviations). They would also violate independence assumptions of the inferential statistical analyzes performed on the data. Therefore, the data reported in Table 1 consists of average performance scores across multiple classification tasks within the same study. For example, the five correlation coefficients from the Eyben et al. study [26] were averaged to yield one multimodal correlation. This resulted in one data point per study. The only exception was the Lin et al. [7] study, which contributed two data points. This is because the two different data sets analyzed in that study were sufficiently unique to warrant separate consideration. It should also be noted that the Caridakis et al. [17] and Castellano et al. [15] studies used the same data set, but considered different models based on different but non-mutually exclusive subsets of the data. The second column (N) in the Table 1 presents the number of data points that were averaged to produce the aggregate scores.

Table 1. Unimodal and Multimodal Classification Accuracies

			Unimodal Performance										
Ref.	**N**	**Metric**	**Face**	**Voice**	**Txt**	**PosGes**	**EEG**	**Gaze**	**Context**	**Event**	**BioS**	**MM**	**MM Effect (%)**
[25] Busso	1	Acc	.851	.709								.891	4.70
[17] Caridakis	1	Acc	.596	.708		.832						.894	7.45
[15] Castellano	1	Acc	.483	.571		.671						.783	16.7
[21] Chanel	1	Acc					.560				.590	.630	6.78
[18] Cueva	1	Acc	.200	.650								.750	15.4
[22] D'Mello	2	Acc	.352			.316			.381			.487	6.83
[11] Emerich	1	Acc	.907	.877								.930	2.54
[26] Eyben	5	CC	.131	.326						.318		.403	-5.71
[13] Glodek	4	Acc	.518	.533								.501	-8.18
[12] Gunes	1	Acc	.829			1.00						.910	-9.00
[19] Gunes	1	Acc	.352			.769						.827	7.54
[8] Jiang	1	Acc	.468	.522								.665	27.4
[23] Kapoor	1	Acc	.668			.820			.572			.865	5.53
[27] Karpouzis	1	Acc	.670	.730								.820	12.3
[4] Kessous	1	Acc	.483	.571		.671						.783	16.7
[14] Khalili	1	Acc					.667				.517	.622	-6.75
[9] Kim	1	Acc		.540							.510	.550	1.85
[7] Lin(a)	1	Acc	.622	.603								.781	25.7
[7] Lin(b)	1	Acc	.714	.710								.906	27.0
[32] Litman	4	Acc		.608	.645							.660	2.63
[10] Metallinou	2	Acc	.562	.559								.630	2.52
[29] Nicolaou	2	CC	.603	.515		.502						.719	10.7
[5] Paleari	1	Acc	.321	.361								.430	19.1
[33] Rabie	1	Acc	.745	.619								.782	4.98
[24] Schuller	3	CC		.683	.685							.776	2.55
[20] Soleymani	2	Acc					.563	.689				.725	5.15
[16] Wagner	1	Acc	.480	.510		.420						.550	7.84
[6] Wollmer	2	Acc	.497	.511								.672	21.5
[31] Wu	1	Acc		.800	.809							.836	3.25
[34] Zeng	1	Acc	.390	.690								.750	8.70

N refers to the number of data points that were aggregated to produce the averages in the table. **Metric.** refers to the metric used to quantity classification accuracy. Acc. = proportion correct. Corr. = correlation coefficient. **PosGes.** refers to posture or gesture. **BioS** refers to a combination of physiological measures (e.g., EMG, ECG) but does not include EEG. **MM.** refers to multimodal performance. **MMEffect (%)** refers to the percent improvement of multimodal affect detection accuracy compared to the best unimodal classifier.

4. RESULTS

4.1 Overall Effects (MM Effect)

The distribution of MM effects is presented in Figure 1. A one-sample t-test indicated that the mean MM effect of 8.12% significantly differed from zero, $t(29) = 4.56$, $p < .001$, $d = .83$ sigma. This suggests that the multimodal classifiers yield non-zero improvements in performance (classification accuracy or correlation) compared to the best unimodal classifiers. There was also considerable variance in the MM effect distribution. MM effects ranged from -9.00% to 27.4% with a standard deviation of 9.75%. The large range and the fact that the standard deviation is greater than the mean, suggests that the *median value of 6.81%* might provide a more accurate estimate of the central tendency of the distribution.

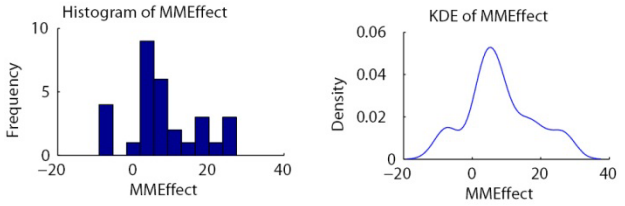

Figure 1. Histogram (left) and kernel smoothing density estimation (right) of distribution of MM effects

4.2 Relationship between Unimodal and Multimodal Accuracies

We investigated the amount of shared variance between the unimodal and multimodal classifiers by correlating accuracy of the best unimodal classifier with multimodal accuracy. This analysis focused on the 27 studies that used classification accuracy as the performance metric. There was a very robust correlation between accuracy of the best unimodal classifier and multimodal accuracy, $r(25) = .901$, $p < .001$. The correlation between multimodal accuracy and accuracy of the second-best unimodal classifier was similarly large, $r(25) = .664$, but was lower than the correlation with the best unimodal classifier. Accuracies of the best and second-best unimodal classifiers were also strongly correlated, $r(25) = .733$, $p < .001$.

To address the extent to which the best and second-best unimodal classifiers explained unique variance in predicting multimodal performance, two partial correlations were computed. First, multimodal accuracy was correlated with the best unimodal accuracy after controlling for the second-best unimodal accuracy. This correlation was statistically significant and was quite large, $r(24) = .814$, $p < .001$. Second, multimodal accuracy was correlated with the second-best unimodal accuracy after partialling out the best unimodal accuracy. This yielded a non-significant correlation, $r(24) = .011$, $p = .957$.Taken together, these results suggests that much of the variance in multimodal accuracy can be explained by accuracies of the best unimodal accuracy. The second-best unimodal classifier did not explain any additional variance.

4.3 Effects as a Function of Data Type

We analyzed how MM effects of classifiers that were trained on naturalistic affective data (natural classifiers) compared to classifiers trained on acted data (acted classifiers). This analysis was complicated by the fact that a majority (16 classifiers) used acted data, but only a handful (5 classifiers) were trained on naturalistic data. To partially address this imbalance, we merged the 5 natural classifiers with the 6 seminatural classifiers to form a new category of 11 *natural-seminatural* classifiers.

An analysis of the MM effect distributions (not shown here) for the acted and natural-seminatural classifiers indicated that each distribution had one potential outlier. These outliers were quantitatively identified as MM effects that exceeded two standard deviations from the mean. There was one outlier (-9.00%) in the distribution of acted MM effects and another (25.7%) in the distribution of natural-seminatural effects. Each outlier was replaced with the next closest value in the distribution. In particular, the -9.00% outlier was replaced with 2.54% and the 25.7% outlier was replaced with 12.3%. Paired-sample t-test on the distributions before and after outlier replacement did not yield significant differences for either the acted ($p = .333$) or the natural-seminatural distributions ($p = .341$), thereby indicating that this method of replacing outliers had no unintended effects.

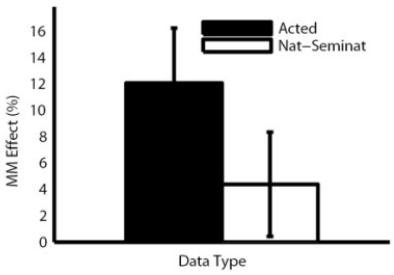

Figure 2. Comparison of data types

Figure 2 shows means and 95% confidence intervals for each effect size distribution. The mean MM effect for the acted classifiers ($M = 12.1$, $SD = 8.51$) was approximately *three times greater* than the mean MM effect for the natural-seminatural classifiers ($M = 4.39$, $SD = 6.70$). An independent samples t-test indicated that this difference was statistically significant, $t(25) = 2.51$, $p = .019$, and was consistent with a large effect ($d = 1.01$ sigma).

We investigated if the MM effects of each data type significantly differed from zero with two one-sample t-tests. A significant difference from zero was obtained for the acted MM effects, $t(15) = 5.68$, $p < .001$, $d = 1.42$ sigma. The test for the natural-seminatural distribution approached significance, $t(10) = 2.17$, $p = .055$, $d = .655$ sigma. This difference is likely to be significant with a larger sample.

4.4 Predicting Multimodal Accuracy

The results so far have indicated that multimodal accuracies were related to accuracies of the best unimodal classifiers (Section 4.2) as well as the type of data (i.e., acted vs. natural-seminatural) used to train the classifiers (Section 4.3). Could these two factors predict MM accuracy with sufficient precision to allow predictions to be made for unseen (new) MM classifiers? This question was addressed by regressing MM accuracy on the best unimodal accuracy and data type (an indicator variable with natural-seminatural as the reference group). The analysis focused on a subset of 24 systems that used classification accuracy as the performance metric and were trained on acted or natural-seminatural data. A tolerance analysis yielded tolerance values of .943, thereby alleviating any multicollinearity concerns pertaining to these two predictor variables [35].

The results yielded a significant model, $F(2, 21) = 54.9$, $p < .001$, that explained a robust amount of the variance ($R^2 = .840$). The dominant predictor was the accuracy of the best unimodal classifier ($\beta = .878$, $p < .001$). Data type was not a significant predictor of MM accuracy ($\beta = .127$, $p = .172$), ostensibly because

the stronger predictor suppressed its effect. This suggests that the two predictor variables did not have an additive effect in predicting MM accuracy.

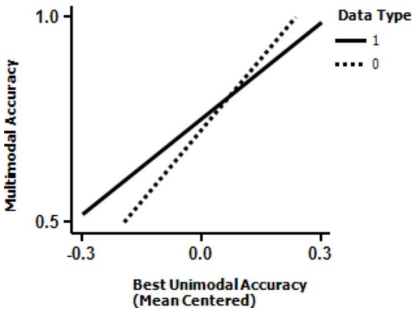

Figure 3. Data Type (1 for acted and 0 for natural-seminatural) × Best Unimodal Accuracy Interaction

We investigated if these two predictors had an interactive effect by including the *data type* × *best unimodal accuracy* interaction term in the regression equation. The overall model was significant, $F(3, 20) = 41.2$, $p < .001$, and yielded an R^2 of .861. The interaction term approached significance ($p = .096$) and explained an additional 2.1% of the variance. This categorical × continuous interaction indicates that the relationship between the best unimodal accuracy and MM accuracy varies by data type (see Figure 3). Specifically, accuracy of the best unimodal classifier was a better predictor of MM accuracy for the natural-seminatural classifiers ($B = 1.17$, $p < .001$) than the acted classifiers ($B = .783$, $p < .001$).

4.5 Generalizability Across Studies

The results of the previous section indicated that it was possible to predict MM accuracy primarily on the basis of the accuracy of the best unimodal classifier. The relatively small sample size raises some important generalizability issues. That is, to what extent does the regression model generalize to new studies? We addressed this question by performing a bootstrapping cross validation analysis. This is the recommended validation technique to assess generalizability and overfitting of models constructed on small data sets [36]. The analysis proceeded as follows. Training data was obtained by sampling a subset of the studies; models were fit on the training data and *training* R^2 was computed. The training models were then applied to the testing data, which consisted of the training studies plus new studies not included in the training data. Goodness of fit for the testing data (*testing* R^2) was obtained. Overfitting was computed as the difference between testing and training R^2 values. This procedure was repeated for 10,000 iterations and average R^2 values were computed. The results yielded an average training R^2 of .826 and an average testing R^2 of .806. The very small discrepancy of .020 suggests that the regression model is likely to generalize to new studies.

4.6 Second Order Effects (MM2 Effect)

Our results, for the most part, have focused on the MM effect distribution or on comparing multimodal accuracy with the best unimodal accuracy. A related question is how multimodal accuracy compares with respect to the *second-best* performing unimodal system. This question was addressed by computing an MM2 effect score according to the following equation: $100 * \frac{a_{12} - max\ 2(a_1, a_2)}{max2\ (a_1, a_2)}$ ($max2$ represents the second largest value in a series). The examination of the distribution of MM2 effects yielded three outliers (92.3%, 135%, and 275%), which were replaced with the next-closest value (62.5%) that was not an outlier.

MM2 effects ranged from 4.23% to 62.5% with a mean of 29.4% (*SD* = 17.6%). The median value of 28.2% was very close to the mean. A one-sample t-test indicated that the mean MM2 effect significantly differed from zero, $t(29) = 9.14$, $p < .001$, $d = .1.67$ sigma. Furthermore, a paired samples t-test indicated that the mean MM2 effect was significantly, $t(29) = 7.05$, $p < .002$, and substantially ($d = 1.29$ sigma) greater than the mean MM effect. Indeed, the mean MM2 effect (29.4%) was approximately 4 times greater than the mean MM effect (8.12%).

We also examined if the type of training data (acted vs. natural-seminatural) influenced the MM2 effects. However, unlike the results for MM effects, an independent samples t-test indicated that the mean MM2 acted effect of 32.9% (*SD* = 19.2) was statistically equivalent to the mean MM2 natural-seminatural effect of 27.2% (*SD* = 16.4%), $t(25) = .809$, $p = .426$.

4.7 Unimodal Comparisons

A long standing question in the literature is whether any one particular modality yields classification accuracies that are superior to the other modalities. This is a difficult question to address in any individual study because it is unclear if any modality advantages obtained in one study will generalize to other studies. Nineteen of the studies monitored facial and acoustic-prosodic features, so there might be sufficient data to draw some generalizations about the comparative advantages of these two modalities. Two of these 19 face-voice studies used the correlation coefficient as the performance metric, while the remaining 17 studies used classification accuracy. This analysis focused on the latter 17 studies.

The results indicated that classification accuracies associated with unimodal facial feature tracking (*M* = .559, *SD* = .182) were quantitatively lower that unimodal voice feature tracking (*M* = .614, *SD* = .118). However, a paired-samples t-test did not yield a significant modality effect, $t(17) = 1.60$, $p = .129$, $d = .39$ sigma. Nevertheless, the .39 effect size in favor of voice is indicative of a small to medium sized effect [37]. This effect might be significant with a larger sample, although this is entirely an empirical question.

As a follow-up analysis, we investigated the pattern of correlations of accuracies obtained by the face, voice, and multimodal systems that included these modalities. Both face and voice accuracies were significantly correlated with multimodal accuracy, but the voice ($r(15) = .870$, $p < .001$) demonstrated a stronger correlation than the face ($r(15) = .614$, $p = .009$). Face and voice accuracies were also related, ($r(15) = .633$, $p = .006$).

5. GENERAL DISCUSSION

The present study analyzed 30 published research articles that developed and validated multimodal affect classifiers. Our focus was on quantifying the improvement that multimodal classifiers afford over their unimodal counterparts. The results were illuminating in a number of respects. In this section, we summarize our major findings, discuss their theoretical implications, address limitations, and identify potential avenues for further research.

5.1 Major Findings and Applied Implications

The major findings are organized into three themes as discussed below.

Significant but modest MM effects. The results consistently revealed that MM effects were significantly greater than zero, with only four of the 30 studies reporting negative effects. This

provides some initial evidence that MM classifiers do outperform their best unimodal counterparts. One caveat is the possibility of publication bias because it is likely that the papers that report positive MM effects are more likely to be published, and subsequently included in this meta-analysis, than papers that report negligible or negative effects.

The results also indicated that MM effects were somewhat modest, with the median overall effect being 6.81% (the mean was 8.12%). Importantly, the effects associated with classifiers trained on naturalistic or seminatural affect data (4.39%) were substantially lower than classifiers trained on acted data (12.1%). Since the ultimate goal of affect detection is to sense naturalistic affective expressions in real-world contexts, the 4.39% effect might represent a more accurate estimate of state-of-the-art multimodal affect detection accuracies. The question of whether this modest improvement in accuracy obtained by MM systems is worth their increased complexity is a question that is best addressed at the application-level.

Redundancy among modalities. One reason for the relatively modest MM effect, especially for the systems trained on more naturalistic data, is that there might be considerable redundancy among the different modalities. Strong correlations among the best unimodal, second-best unimodal, and MM accuracies provide some evidence to support this view. Further evidence for redundancy among modalities can be obtained by the fact that the best unimodal accuracies predicted a cross-validated 80.6% of the variance in multimodal accuracies. Indeed, impressive MM effects are not expected if the different modalities convey similar information, albeit in different ways.

Importantly, accuracy of the best unimodal classifier was a better predictor of MM accuracy for the natural-seminatural systems compared to the acted systems (see Section 4.4). This finding suggests that natural-seminatural MM systems have less room for improvement than acted systems. This finding is intuitively plausible because individuals tend to invoke a prototypical emotional response when asked to "act out" an emotion. This typically involves a higher level of coordination among the different modalities when compared to naturalistic expressions of emotion, thereby resulting in more optimistic MM effects.

Substantial second-order effects. The analysis that focused on assessing MM performance improvements over the second-best unimodal classifier yielded particularly interesting findings. The mean MM2 effect was an impressive 29.4% and was not dependent upon whether the training data was acted or natural-seminatural. Additionally, a paired-samples t-test indicated that the accuracy of the second-best unimodal system (M = .559, SD = .161) was significantly ($t(26)$ = -5.30, $p < .001$) and substantially (d = 1.02 sigma) lower than accuracy of the best unimodal system (M = .674, SD = .143). Taken together, these findings suggest that although combining modalities yields modest improvements in affect detection accuracies, *considering multiple individual modalities* can have a major impact on system performance. This is because performance would be severely impacted if only one modality was modeled and in the worst case if it always happened to be the lower performing modality.

5.2 Theoretical Implications

The fact that combining multimodal accuracies yielded modest improvements has important implications for psychological theories of emotion. These theories in turn guide much of our 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, Ekman, Izard, and others, posits that discrete "affect programs" produce the physiological, behavioral, and subjective changes associated with a particular emotion [38-40]. 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 multimodal affect detection should yield accuracies that are substantially greater than the individual modalities 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 [41, 42]. According to this 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. Therefore, coordinated bodily responses associated with particular emotions are rare. These models would accommodate the prediction that in most cases a combination of modalities might conceivably yield small improvements in classification accuracies.

We suspect that the expectation for impressive multimodal effects stems from an adherence to the classical model of emotion. However, the present data is more consistent with the alternate approach, which suggests that other than the rare cases of prototypical emotions, or in artificial experimental contexts involving acted emotions, modest MM effects might be expected.

5.3 Limitations and Future Work

There are two primary limitations to this work. The first pertains to the comprehensiveness of the studies that were analyzed. Our focus was on obtaining a reasonably large sample of studies that reported MM accuracies, rather than attempting to analyze every single study in the literature. This is defendable because one does not need to study an entire population to estimate its parameters. Furthermore, almost all of the tests of statistical significance yielded significant results, thereby suggesting that our sample size was adequate to detect the relatively large effects in our data.

The second limitation was that there was some imbalance with respect to the modalities, data, evaluation metrics, and affective states classified. For example, a majority of the studies we analyzed focused on audio-visual affect recognition, so the results are somewhat biased towards these systems. It is important to note, however, that this imbalance in our study is linked to a similar imbalance in the current state-of-the art. Specifically, most studies focus on the audio and visual modalities, while EEG, gaze, and context-based sensing are comparatively rare. Physiological-based affect sensing (i.e., biosignals) are quite popular affect detection modalities, but these are not often combined with face, voice, text, and other modalities.

This form of data imbalance was also the reason why we did not perform moderation analyses on other study-level variables (e.g.,

affective states classified, specific fusion methods) like we did with data type.

We are in the process of addressing these limitations in two ways. First, we are expanding the analysis to include a larger number of studies (approximately 50-100). Second, we will increase the breadth of our search of the literature to include all available MM studies on some of the less common modalities.

5.4 Concluding Remarks

The phrase "*consistent, but modest*" succinctly captures the results of this study. These MM systems were *consistently* better than their unimodal counterparts, but the improvements were *modest*, at least for the natural-seminatural systems. A fundamental question is whether these findings can be best explained by the *method* or by the *data.* In particular, are MM effects modest because our classifiers are not sufficiently sophisticated to model the intricate nonlinear time-varied relationships between the different modalities? Or are they modest because the data used to train the classifiers does not contain adequate expressions of coordination and synchronization among modalities, thereby rendering even the most sophisticated classifiers inept? The field of multimodal affect detection is too young to currently settle these issues, so the answer to this question awaits further research.

However, there is another possibility beyond the method and the data. It may be the case that the expression of naturalistic emotions is inherently a diffuse phenomenon, which will usually yield modest effects irrespective of method or data. This suggests that in addition to considering different methods and data sources, it might be useful to consider alternate models of emotion beyond the classic view described in Section 5.2. Unfortunately, almost all of the 30 studies (including our own [22]) we analyzed emphasized the method and the data at the expense of examining the affective phenomenon itself (i.e., insufficient attention to recent development in emotion theories and alternate models). Perhaps a more balanced approach that combines better data sources and innovative classifiers with more diverse emotion models represents the most promising way forward.

6. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (NSF) (ITR 0325428, HCC 0834847, DRL 1235958). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

7. REFERENCES

[1] Calvo, R.A. and D'Mello, S.K. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1, 18-37.

[2] Zeng, Z., Pantic, M., Roisman, G. and Huang, T. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31, 39-58.

[3] Pantic, M. and Rothkrantz, L. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91, 1370-1390.

[4] Kessous, L., Castellano, G. and Caridakis, G. 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3, 33-48.

[5] Paleari, M., Benmokhtar, R. and Huet, B. 2009. Evidence theory-based multimodal emotion recognition. In *Proceedings of Proceedings of the 15th International Multimedia Modeling Conference (MMM '09)* (Chongqing, China, January 6-8, 2009).Springer-Verlag, Berlin, Heidelberg, 435-446.

[6] Wöllmer, M., Metallinou, A., Eyben, F., Schuller, B. and Narayanan, S.S. 2010. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In *Proceedings of Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)* (Makuhari, Chiba, Japan, September 26-30, 2010). 2362-2365.

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

[8] Jiang, D., Cui, Y., Zhang, X., Fan, P., Ganzalez, I. and Sahli, H. 2011. Audio visual emotion recognition based on triple-stream dynamic bayesian network models. In *Proceedings of Fourth International Conference on Affective Computing and Intelligent Interaction* (Memphis TN, October 9-12, 2011).Springer-Verlag, Berlin Heidelberg, 609-618.

[9] Kim, J. 2007. Bimodal emotion recognition using speech and physiological changes. In Grimm, M. and Kroschel, K. eds. *Robust Speech Recognition and Understanding*, Vienna, Austria, I-Tech, 265-280.

[10] Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B. and Narayanan, S. in press. Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification. *IEEE Transactions on Affective Computing*.

[11] Emerich, S., Lupu, E. and Apatean, A. 2009. Emotions recognition by speech and facial expressions analysis. In *Proceedings of Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009)* (Glasgow, Scotland, August 24-28, 2009).

[12] Gunes, H. and Piccardi, M. 2005. Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In *Proceedings of First International Conference on Affective Computing and Intelligent Interaction (ACII 2005)* (Beijing, China, October 22-24, 2005).Springer-Verlag, Berlin Heidelberg, 102-111.

[13] Glodek, M., Tschechne, S., Layher, G., Schels, M., Brosch, T., Scherer, S., Kächele, M., Schmidt, M., Neumann, H. and Palm, G. 2011. Multiple Classifier Systems for the Classification of Audio-Visual Emotional States. In *Proceedings of 4th International Conference on Affective Computing and Intelligent Interaction (ACII 2011)* (Memphis, TN, October 9-12, 2011).Springer, Berlin/Heidelberg, 359-368.

[14] Khalali, Z. and Moradi, M. 2009. Emotion recognition system using brain and peripheral signals: Using correlation dimension to improve the results of EEG. In *Proceedings of Proceedings of International Joint Conference on Neural Networks* (Atlanta, GA, June 14-19, 2009).IEEE, 1571 - 1575

[15] Castellano, G., Kessous, L. and Caridakis, G. 2008. Emotion recognition through multiple modalities: face, body gesture, speech. In Peter, C. and Beale, R. eds. *Affect and Emotion in*

Human-Computer Interaction (LNCS, vol. 4868), Springer, Heidelberg, 92-103.

[16] Wagner, J., Andre, E., Lingenfelser, F., Kim, J. and Vogt, T. 2011. Exploring Fusion Methods for Multimodal Emotion Recognition with Missing Data. *IEEE Transactions on Affective Computing*, 2, 206-218.

[17] Caridakis, G., Castellano, G., Kessous, L., Raouzaiou, A., Malatesta, L., Asteriadis, S. and Karpouzis, K. 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In Boukis, C., Pnevmatikakis, L. and Polymenakos, L. eds. *International Federation for Information Processing, Volume 247, Artificial Intelligence and Innovations 2007: From Theory to Applications*, Springer, Boston, 375-388.

[18] Cueva, D., Gonçalves, R., Cozman, F. and Pereira-Barretto, M. 2011. Crawling to Improve Multimodal Emotion Detection. In *Proceedings of Proceedings of the 10th Mexican International conference on Artificial Intelligence (MICAI 2011)* (Puebla, Mexico, November 26 - December 4, 2011).Springer-Verlag, Berlin, Heidelberg, 343-350.

[19] Gunes, H. and Piccardi, M. 2009. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 39, 64-84.

[20] Soleymani, M., Pantic, M. and Pun, T. in press. Multi-Modal Emotion Recognition in Response to Videos. *IEEE Transactions on Affective Computing*.

[21] Chanel, G., Rebetez, C., Bétrancourt, M. and Pun, T. 2011. Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 41, 1052-1063.

[22] D'Mello, S. and Graesser, A. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-adapted Interaction*, 20, 147-187.

[23] Kapoor, A. and Picard, R. 2005. Multimodal affect recognition in learning environments. In *Proceedings of Proceedings of the 13th annual ACM international conference on Multimedia* (Hilton, Singapore, November 6-11, 2005).ACM, NY, NY, 677-682.

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

[25] Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C.M., Kazemzadeh, A., Lee, S., Neumann, U. and Narayanan, S. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of Proceedings of the 6th International Conference on Mutlmodal Interfaces (ICMI '04)* (State College, PA, October 13-15, 2004).ACM, New York, NY, 205-211.

[26] Eyben, F., Wollmer, M., Valstar, M.F., Gunes, H., Schuller, B. and Pantic, M. 2011. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In *Proceedings of Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)* (Santa Barbara, CA, March 21-25, 2011).IEEE, Washington, DC, 322-329.

[27] Karpouzis, K., Caridakis, G., Kessous, L., Amir, N., Raouzaiou, A., Malatesta, L. and Kollias, S. 2007. Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition. In Huang, T. ed. *Artifical Intelligence for Human Computing*, Springer-Verlag, Berlin Heidelberg, 91-112.

[28] McKeown, G., Valstar, M., Cowie, R., Pantic, M. and Schroder, M. 2012. The SEMAINE database: Annotated multimodal records of emotionally coloured conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3, 5-17.

[29] Nicolaou, M., Gunes, H. and Pantic, M. 2011. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence& Arousal Space. *IEEE Transactions on Affective Computing*, 2, 92-105.

[30] Adolphs, R.A. 2002. Neural systems for recognizing emotions. *Current Opinion in Neurobiology*, 12, 169-177.

[31] Wu, C. and Liang, W. 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2, 10-21.

[32] Litman, D.J. and Forbes-Riley, K. 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48, 559-590.

[33] Rabie, A., Wrede, B., Vogt, T. and Hanheide, M. 2009. Evaluation and discussion of multi-modal emotion recognition. In *Proceedings of Proceedings of the Second International Conference on Computer and Electrical Engineering (ICCEE '09)* (Dubai, UAE, December 28-30, 2009).IEEE Computer Society, 598-602.

[34] Zeng, Z., Hu, Y., Roisman, G., Wen, Z., Fu, Y. and Huang, T. 2006. Audio-visual emotion recognition in adult attachment interview. In *Proceedings of Proceedings of the 8th international conference on Multimodal Interfaces (ICMI '06)* (Banff, Canada, November 2-4, 2006).ACM, New York, NY, 139-145.

[35] Allison, P.D. 1999. *Multiple regression*. Pine Forge Press, Thousand Oaks, CA.

[36] Baayen, R.H. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge Univ Press, Cambridge.

[37] Cohen, J. 1992. A power primer. *Psychological Bulletin*, 112, 155-159.

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

[39] Tomkins, S.S. 1962. *Affect Imagery Consciousness: Volume I, The Positive Affects*. Tavistock, London.

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

[41] Coan, J.A. 2010. Emergent ghosts of the emotion machine. *Emotion Review*, 2, 274-285.

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