Exploring the Contextual Factors Affecting Multimodal Emotion Recognition in Videos

Prasanta Bhattacharya [ID], Raj Kumar Gupta [ID], and Yinping Yang [ID]

Abstract—Emotional expressions form a key part of user behavior on today's digital platforms. While multimodal emotion recognition techniques are gaining research attention, there is a lack of deeper understanding on how visual and non-visual features can be used to better recognize emotions in certain contexts, but not others. This study analyzes the interplay between the effects of multimodal emotion features derived from facial expressions, tone and text in conjunction with two key contextual factors: i) gender of the speaker, and ii) duration of the emotional episode. Using a large public dataset of 2,176 manually annotated YouTube videos, we found that while multimodal features consistently outperformed bimodal and unimodal features, their performance varied significantly across different emotions, gender and duration contexts. Multimodal features performed particularly better for male speakers in recognizing most emotions. Furthermore, multimodal features performed particularly better for shorter than for longer videos in recognizing neutral and happiness, but not sadness and anger. These findings offer new insights towards the development of more context-aware emotion recognition and empathetic systems.

Index Terms—Affective computing, affect sensing and analysis, modelling human emotions, multi-modal recognition, sentiment analysis, technology & devices for affective computing

✦

1 INTRODUCTION

FOR long, the interest surrounding emotion recognition has been a key focus area within the field of affective computing [1], [2], [3]. Recent advances in computing infrastructure and datasets have led to new developments in detecting emotions using multiple types of data modalities like facial and acoustic expressions, linguistic and semantic patterns, body movements, eye gaze patterns and electroencephalography signals [2]. Recent studies have proposed different design approaches to perform automatic recognition of affective outcomes like valence, arousal, dominance, and emotion types [1], [2].

The ability to automatically recognize emotions has valuable implications for a range of applications. These include the design of empathetic agents [4], understanding consumer behavior at scale from observational data [5], and aiding healthcare professionals in a range of activities from diagnosing depressive symptoms in patients, to tracking the onset and progression of autism [6], [7].

This paper focuses on extracting high-level emotion features from visual, audio and text data modalities by leveraging a number of recent emotion analysis technologies. We tested the efficacy of these features in classifying fine-grained emotions from a real-world dataset comprising 2176 labelled video clips, and sought to explore the

conditions under which multimodal features outperformed bimodal and unimodal features.

More importantly, we performed an in-depth analysis of the effects from two important contextual factors that impact the performance of multimodal emotion recognition systems: i) *gender* of the speaker, and ii) *duration* of the emotion expression episode. First, we investigated whether the speaker's gender was linked to the relative effectiveness of visual, audio and text features in recognizing different emotions. To date, despite extensive research aimed at uncovering gender differences in how emotion is experienced and perceived (e.g., McDuff *et al.* [8]), little is known about how these differences affect the predictive performance of affective computing systems. We found that while multimodal features tend to outperform unimodal features for both male and female speakers, the relative improvement of using multimodal over unimodal was higher for male speakers. A trimodal classifier (i.e., combining visual, audio, and text features) was found to be the best performing multimodal classifier for both genders. For the unimodal features alone, we found that visual features performed best for both male and female speakers. With respect to the emotion categories in the overall sample, we found that while visual features performed best for predicting happy, sad, and neutral emotions, the audio features performed best for anger. Through these insights, the current study aims to be among the first to emphasize the importance of developing gender-aware multimodal systems, as the empirical performance of these systems can be largely contingent on inherent gender differences among speakers.

In addition to gender, we examined if the duration of the emotion being expressed affected the way a particular emotion was perceived and recognized across various modalities. Psychological studies in the past have examined the

• The authors are with the Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), Singapore 138632. E-mail: {prasanta_bhattacharya, gupta-rk, yangyp}@ihpc.a-star.edu.sg.

relationship between emotions and emotion duration using subjects' recalled emotional experiences [9], [10], [11]. However, to the best of our knowledge, the relationship between episodic duration of emotions and classifier performance in videos, as well as the associated underpinnings, are still not well understood. Our analyses show that duration played an important role in affecting classifier performance. Specifically, while multimodal classifiers outperformed unimodal classifiers for videos of all durations, this relative improvement was higher in shorter episodic durations for only the negative emotions like sadness and anger. In contrast, the relative improvement from using multiple modalities was higher in longer episodic durations for neutral and happiness emotions.

2 CONCEPTS AND RELATED WORK

Expressing and perceiving emotions is central to human experience and is a key contributor to the sustenance of interpersonal communication [12], [13], [14]. Emotions also affect how we interpret our environment, develop opinions and form judgments about the surrounding individuals and situations, and even drive behaviors [15].

Within the overarching category of 'affect', it is generally accepted that 'emotion' is distinct from related states like mood and temperament. Emotions might involve a range of feelings which may be transient, like relief, or last for relatively long periods of time, as is the case with mood, or really long spans of time, as with temperament [16], [17], [18]. Some of the earliest definitions about emotions came independently from psychologist William James and physiologist Carl Lange who contended that emotions arise naturally as a result of environmental stimulus, and that our interpretation of the physiological sensation of these changes constitutes emotion [19]. Contrary to James' contention that emotions could be a response to environmental changes, W.B. Cannon treated emotions as being felt first but expressed outwardly in certain behaviors much later [20]. There is also debate over how emotions can be perceived or described [13], [14], [21].

2.1 Emotion Type, Valence, Arousal, and Intensity

For the purpose of this paper, we adopt three pertinent concepts from the emotion literature to facilitate the subsequent emotion feature extraction and analysis.

First, we refer to the stream of work on *categorization* of emotions which views emotions as discrete and distinct types such as happiness (or joy), anger, sadness and fear [22], [23], [24], [25], [26]. This view, while "simplistic"[27], [28], helps to focus the present study in terms of differentiating the specific effects due to data modality, duration and gender for various distinct emotion types. Second, we leverage the *dimensional* view of emotion that has received growing popularity in recent years [29], [30], [31], [32], [33]. In this study, we specifically considered two dimensions: valence and arousal. *Valence* describes the degree of pleasantness or unpleasantness of a signal and is generally measured on a continuous scale ranging from positive (i.e., pleasant) to negative (i.e., unpleasant). *Arousal*, on the other hand, refers to the extent of physical activation, ranging from no arousal to very high arousal. Thus, various

emotional states like happy, angry, sad and afraid can be mapped to a two-dimensional valence-arousal scale, known as the circumplex model of affect [34]. For instance, in the circumplex model [34], "happy" is associated with a positive valence and moderate to high arousal. In contrast, both "tensed" and "angry" have a negative valence with moderate to high arousal, while "sad" is associated with a negative valence and low arousal. Third, we also consider emotion *intensity* as an important concept. This aspect of emotion concerns the degree or "depth" of emotional experience, ranging from barely noticeable to most intense imaginable [26], [35], [36]. For instance, the anger emotions can range from low-intensity annoyance to high-intensity rage, and this difference represents a relatively new, less explored aspect of emotional information that can benefit the development of computational methods (e.g., [37]).

2.2 Multimodal vs Unimodal Recognition Systems

Early research into emotion recognition systems largely focused on single data modalities (e.g., only facial expressions or tone of voice) and on emotions extracted from enacted sequences by trained actors [38]. More recently, studies have attempted to combine signals from multiple modalities such as facial expressions and audio [39], [40], audio and written text [41], [42], physiological signals [43], and various combinations of these modalities [44]. Overall, studies in this field have started to emphasize the importance of multimodal signals of emotions in more naturalistic scenarios (See [2], [3] for detailed reviews).

However, despite notable efforts, the existing literature on emotion recognition is limited in three ways. First, while there is growing evidence that multimodal systems generally outperform unimodal systems in relevant tasks [2], [40], [45], recent results have shown more complex patterns. For example, a number of studies have found that multimodal systems can often exhibit negligible improvements or even reductions in performance [39], [46]. Second, a vast majority of existing analyses on multimodal systems were based on relatively small samples (i.e., fewer than 50 participants), and using mostly bimodal classifiers. Third, there is also scant work on understanding the role of various kinds of contextual factors that might directly or indirectly influence model performance (e.g., de Gelder *et al.* [47]). With the exception of a few studies, the extant literature does not offer an adequate understanding of why certain data modalities work better for specific emotions, but not others [48]. It is therefore important to address the fundamental question of which data modalities are best suited for certain emotions, and the role of contextual factors in affecting the performance of these multimodal systems.

2.3 Contextual Factors: Gender and Duration

How does the performance of multimodal emotion recognition systems change in different contexts? A few recent studies, such as [39], [47], [49], [50], have started to explore how contextual factors can potentially affect the performance of emotion related classifiers. For instance, de Gelder *et al.* investigated how the presence of information on the visual and auditory context (e.g., information about natural scene, vocal expressions etc.) can benefit the recognition of facial

expressions [47]. Contextual information can also come from the same modality. For instance, Metallinou *et al.* showed that emotions within as well as across utterances in a dialog sequence can be leveraged to improve emotion recognition [39]. These studies have focused primarily on *external* contexts (e.g., speaker's environment, emotion eliciting factors).

A relatively new direction is the interplay of *internal* contexts manifested through individual differences, like gender, and emotions. For example, Brebner analyzed two samples comprising Australian and International participants for differences in the frequency and intensity of self-reported emotions by gender [51]. The results highlighted significant differences for affection and sadness emotions, where females scored higher on both frequency and intensity. Males, on the other hand, scored higher on pride in both frequency and intensity. This is not entirely surprising, and a number of studies have attempted to explain gender differences in how emotions are experienced and perceived in others [8], [52], [53]. For instance, it has been shown that females generally perform better than men at perceiving and exhibiting negative emotions like sadness and fear [54]. The reasons for this range from evolutionary (e.g., Babchuck *et al.*'s primary caretaker hypothesis or Hampson *et al.*'s fitness threat hypothesis [52], [55]) to biological [54] to social and normative factors [56]. However, it remains to be understood whether and how the performance of multimodal emotion features varies as a result of such gender differences.

Another contextual factor that has attracted recent attention, and motivates the present study, is the role of *temporal* contexts. Recent studies have shown that the temporal length of the specific emotion being recorded is likely to influence the effectiveness of the classifier [48], [49]. This is likely due to a combination of three major factors. First, the duration of an emotion expression episode is likely to vary as a function of the data modality [49]. For example, Lingenfesler *et al.* [48] showed that an asynchronous fusion algorithm that takes into account different onset and offset times for emotions in various modalities, outperformed synchronous fusion algorithms. This is particularly useful since certain modalities, like language, spoken or written, take longer to manifest emotions, while others like facial expressions, take shorter time. Second, the nature of the platform or device through which the emotions are being recorded often influences the intensity and duration of the emotion. For instance, users recording a vlog on YouTube are likely to express emotions over longer durations, as compared to users uploading a much shorter Vine video. Lastly, the specific types of emotion might also be associated with the duration of the episode. In this study, we investigate how the duration of the videos affects the performance of multimodal features across various emotions.

3 DATA

3.1 Data Source

We make use of the One-Minute-Gradual (OMG) Emotion Dataset [57] for our analyses. The dataset was released as part of the 2018 OMG Emotional Behavior Challenge[1] where

Fig.1. OMG-emotion dataset (left to right, top to bottom: Actor/actress demonstrating happiness, sadness, anger, disgust, fear, and surprise).

TABLE 1
Descriptive Statistics of the Dataset

#, count of videos
%, out of the total of 2176 videos

Neutral	Happiness	Sadness	Anger	Disgust	Fear	Surprise
772	652	315	260	100	53	24
35.5%	*30.0%*	*14.5%*	*12.0%*	*4.6%*	*2.4%*	*1.1%*

the original task was to implement an emotion-recognition system to accurately predict arousal and valence scores using an annotated dataset. The original OMG train dataset[2] contains YouTube links of 2444 public video clips. We were able to retrieve 2176 clips as of May 2018 using the links provided, as some links became unavailable over time. Each video clip's duration ranged around the one-minute mark in general, but with some shorter and longer duration videos. Most of the videos were of actors and actresses who were practicing lines or monologues probably in preparation for auditions. Through these monologues, the actors and actresses exhibited a number of discrete emotions (e.g., happiness, sadness, surprise), of varying arousals. Fig. 1 shows screenshots of six sample videos from this dataset.

3.2 Emotion Ground Truth Labels

The dataset also provides utterance-level (approximately 10 seconds per utterance) majority-voted ground truth labels ("EmotionMaxVote") for seven discrete emotion categories, namely anger, fear, happiness, neutral, sadness, disgust, and surprise[3]. For each clip, the majority vote label was calculated by having five different annotations from crowd-sourced Amazon Mechanical Turk workers [57].

Table 1 presents the basic descriptive statistics of emotion categories, sorted by the most balanced to the most imbalanced classes. We note that neutral and happiness had acceptable levels of imbalance, while sadness and anger exhibited higher class imbalance. Disgust, fear and surprise classes had the highest class imbalance. The relatively small number of disgust, fear and surprise cases in the dataset is an important limitation. It is for this reason that we choose

to perform the subgroup analyses using only the neutral, happiness, sadness, and anger emotion categories.

Here, it is useful to note that the class labels refer to the emotion ground truth of the speakers as *perceived* by the multiple independent annotators. It is different from *self-reported* or *recalled* emotional experiences, as used in traditional psychological studies.

3.3 Gender and Duration Labels

For this research, we manually annotated the gender of the speaker. We coded a discrete *gender* variable with three categories: "female", "male", and "others", for when it was either not possible to assess the gender or when there were multiple speakers in one utterance. We calculated the *duration* of each video using the start and end time of the video clips mentioned in the dataset.

4 EMOTION FEATURE EXTRACTION

As the purpose of our study is to understand the effects of contextual factors affecting multimodal emotion recognition, our feature extraction strategy focuses on extracting *theoretically explainable, high-level* emotion features from the visual, audio and text channels, respectively. Hence, even though it was possible to extract a large number of lower-level visual and non-visual features (e.g., facial action units, word embeddings), we chose to leverage a set of pre-trained emotion analysis systems that are grounded in data and key emotion concepts (see Section 2.1) to generate high-level emotion features.

4.1 Visual-Based Emotion Feature Extraction

Among all modalities, facial expressions (e.g., frown, smiling) provide arguably the most intuitive emotion cues. We used a face emotion analysis system called *FEA*[4] to extract the visual features. FEA was trained with visual profiling techniques from static images [58], and was later extended to process a set of image frames in videos using deep learning techniques (see details in [58], [59], [60], [61], [62]). The system has been tested on AffectNet test set [63], and achieved Concordance Correlation Coefficients (CCC) of 0.527 and 0.569 for valence and arousal prediction respectively.

For each video, the FEA system generates *vValence* and *vArousal*, which indicate the degree of valence (from very unpleasant to very pleasant) and the degree of arousal (from no to very high physical activation) detected from the speaker's facial expressions. A third feature, *vIntensity,* measures the distance to the neutral state, calculated as the square root of the sum of squares of *vValence* and *vArousal*. The outputs also include 25 binary features (1: yes; 0: no) indicating whether, or not, a face is expressing a particular emotional state aligned with the circumplex emotion model [34]. Taken together, we extracted a total of 28 visual-based emotion features, namely *vValence, vArousal, vIntensity, vAfraid, vAlarmed, vAnnoyed, vAroused, vAstonished, vBored, vCalm, vContent, vDelighted, vDepressed, vDistressed, vDroopy, vExcited, vFrustrated, vGloomy, vHappy, vMiserable, vNeutral, vPleased, vSad, vSatisfied, vSerene, vSleepy, vTensed,* and *vTired.*

4.2 Audio-Based Emotion Feature Extraction

In addition to visual features, the acoustic patterns of speech (e.g., speaking speed, pitch, intonation) provide different sources of emotion cues. For pre-processing, we first applied *FFMPEG*[5] to extract the audio component from the videos. We then extracted the audio-based emotion features using the *AcousEmo*[6] *system*, which was trained with acoustic analysis of emotion signals using deep Boltzmann machines (DBM) ([64], [65], [66], [67], [68], [69]). The DBM classifier showed an effective predictive accuracy, measured as unweighted average recall, of 53.6, 64.4, 56.3 and 53.1 over valence, arousal, power and expectancy dimensions, respectively, when evaluated using the 2011 Audio/Visual Emotion Challenge dataset [70] (see details described in [69]).

For a given audio, AcousEmo produces two main outputs: *aValence* and *aArousal*, indicating the degree of valence (from very unpleasant to very pleasant) and the degree of arousal (from no to very high physical activation) detected from the acoustic signals. The system also generates two additional emotion dimensions: *aPower* (which subsumes two related concepts of the power and control that a speaker is expressing) and *aExpectancy* (which reflects whether the speaker is anticipating or expecting things to happen) (see theoretical discussion of the two additional dimensions [71] and the AVEC2011 challenge [70] for details). For the present OMG data, we extracted a total of ten audio-based emotion features: *aValence, aArousal, aPower, aExpectancy, aIntensity, aJoy, aAnger, aFear, aSadness, aNeutral.*

4.3 Text-Based Emotion Feature Extraction

In addition to the visual and acoustic features, the language content of a speaker's speech in a video (e.g., word choice, linguistic meaning) can carry valuable emotion cues. Before extracting the text-based emotion features, we first obtained the speech transcript from the audio using an Automatic Speech Recognition (ASR) engine developed by I2R[7]. This engine incorporates a range of speech processing techniques including subharmonic ratio based voice activity detection [72], mismatched crowdsourcing for acoustic modeling [73], multi-task learning for pronunciation modeling [74], and multi-task adversarial training for unsupervised adaptation [75], [76]. It has exhibited robust performance at various speech processing tasks including transcribing, retrieving, and indexing speech in English, Mandarin, Malay, Tamil, and other Southeast Asian spoken languages [77], [78], [79], [80]. The base engine we used is trained using a deep neural network (DNN), based on Kaldi speech recognition toolkit. Tested on dev-clean and test-clean conditions[8] of LibriSpeech data, the system achieved 9.32 and 9.49 percent word error rate (WER), respectively.

Next, we adopted *CrystalFeel*[9] to extract the emotion intensity features from the speech transcript. CrystalFeel is a collection of five SVM-based algorithms independently trained with tweets labelled with intensity scores for the overall valence, joy, anger, sadness and fear, respectively

---

4. FEA is accessible via the authors' [58] page https://sites.google.com/site/vonikakis/software-code/appealing_slideshows or via https://opsis.sg (commercial version).

5. FFMPEG is accessed via https://github.com/FFmpeg/FFmpeg
6. AcousEmo is accessible by contacting the authors [65].
7. The I2R ASR engine is accessible by contacting the authors [75].
8. Kaldi training under different conditions: https://github.com/kaldi-asr/kaldi/blob/master/egs/librispeech/s5/RESULTS
9. CrystalFeel is accessible via the authors' [37] page https://socialanalyticsplus.net/crystalfeel.
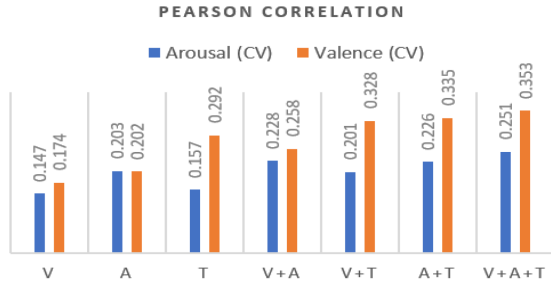
Fig. 2. Pearson correlation coefficients for cross-validated (CV) models.

[37]. Evaluated on the SemEval-18 affect in tweets dataset [81], CrystalFeel achieved high Pearson correlation coefficients of 0.816 (overall valence), 0.708 (joy), 0.740 (anger), 0.700 (fear) and 0.720 (sadness) with human labelled emotion intensity scores [37]. CrystalFeel has also been useful in generating effective emotion features for predicting and understanding happiness in crowdsourced text [82] and predicting news popularity on Facebook [83].

For a given text such as a tweet or a speech transcript, CrystalFeel generates text-based emotion features in terms of the *intensity* value (from 0: barely noticeable to 1: extremely intense) of the five analytic dimensions. We were able to extract and use the following five text-based emotion intensity features from the speech transcript for each video: *tValence, tJoy, tAnger, tFear, tSadness*.

Details on the individual feature description and a bivariate correlational analysis, showing to what extent each feature is correlated with each of the OMG ground truth emotion labels, are provided in Appendix A, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TAFFC.2021.3071503.

5 EXPERIMENTS AND ANALYSES

In order to analyze and compare the effectiveness of different feature sets, we trained a support vector machine (SVM) with a polynomial kernel[10] and tuned the cost function using a grid search approach over the range of [0.01, 10]. This choice of classifier is consistent with recent work in this field (see [84] for an example).

For the classification models (i.e., classifying fine-grained emotion classes), the relevant accuracy measures considered were the raw accuracy and the area under curve (AUC). For the regression models (i.e., predicting arousal and valence scores), the relevant accuracy measure considered was the Pearson correlation coefficient.

5.1 Modality Effects: Comparing Unimodal, Bimodal, and Trimodal Features

We first tested our multimodal features on the task of predicting continuous-valued arousal and valence scores provided by the OMG dataset. We developed a SVM regression model with a linear kernel and grid-search tuned cost function. Fig. 2 demonstrates the Pearson correlation coefficient from the regression models implemented for various feature combinations.

10. The subgroup analyses were performed using a linear kernel.

TABLE 2.
Raw Accuracy and AUC Results for Emotion Classifier Based on 10-Fold Cross-Validation

Modality / Emotion	Neutral	Happiness	Sadness	Anger	Disgust	Fear	Surprise
	Raw Accuracy (%)						
Visual	**57.3**	**71.6**	75.8	61.8	56.8	53.1	73.9
Audio	50.9	54.2	68.5	79.0	77.5	**76.8**	86.0
Text	42.1	66.6	**84.3**	79.5	85.3	43.1	44.4
V + A	**62.0**	**74.9**	76.2	76.5	82.6	72.9	89.6
V + T	59.5	71.1	77.8	65.6	72.9	63.0	78.6
A + T	56.6	59.1	**78.2**	79.7	89.2	86.4	92.5
V + A + T	64.5	75.6	82.4	80.9	88.3	84.1	95.6
	AUC						
Visual	**.583**	**.639**	**.674**	.608	.593	.649	.724
Audio	.569	.598	.635	**.651**	**.673**	**.725**	**.785**
Text	.536	.598	.599	.571	.581	.598	.616
V + A	**.660**	**.695**	**.735**	.720	.790	.842	.927
V + T	.624	.665	.724	.661	.758	.746	.851
A + T	.631	.642	.714	.703	.758	.802	.880
V + A + T	**.688**	**.714**	**.799**	.783	.872	.900	.937

The results show that the trimodal model, denoted by V+A+T, performed the best for both valence and arousal. Specifically, the best multimodal model outperformed the best unimodal model in Pearson correlation scores by over 23 and 20 percent for arousal and valence respectively. Interestingly, while the V+A model was found to be the most effective bimodal model in predicting arousal, the V+T and A+T models performed better in predicting valence. Similarly, while the audio-only (A) model was the best performing unimodal model in predicting arousal, the text-only (T) model performed best in predicting valence.

We then evaluated the multimodal features in classifying the seven emotion categories using a tuned binary SVM, and a 10-fold cross-validation. We applied class weights proportional to the class distribution within each fold to account for the inherent class imbalance.

Table 2 presents the raw accuracy and AUC scores from this analysis. Our results show that the bimodal and trimodal classifiers exhibited significant improvement in raw accuracy and AUC scores, over the unimodal classifiers to varying extents, across the different emotions. We tested the statistical significance of these improvements using a combination of a paired t-test as well as a Wilcoxon signed-rank test, and the differences were statistically significant at the 95 percent confidence level.

To better quantify the improvements in accuracy, we use a version of the commonly used MM1 score which is specified as follows:

$$\text{MM1} = 100 * [Max(a_{MM}) - Max(a_{UM})]/Max(a_{UM}) \qquad (1)$$

Here, $Max(a_{MM})$ denotes the accuracy of the best multimodal classifier ($A+T$ or $V+T$ or $V+A$ or $V+A+T$), $Max(a_{UM})$ denotes the accuracy of the best unimodal classifier (A or V or T). The effect size of the MM1 score has been used in a number of recent studies and offers a conservative

TABLE 3
MM1 Scores Based on 10-Fold Cross-Validated AUC Scores

Modality	MM1 Score (%)						
	Neutral	Happiness	Sadness	Anger	Disgust	Fear	Surprise
	18.0	11.7	18.5	20.3	29.6	24.1	19.4
Best MM	$V+A+T$	$V+A+T$	$V+A+T$	$V+A+T$	$V+A+T$	$V+A+T$	$V+A+T$
Best UM	V	V	V	A	A	A	A

estimate of the relative improvements from using a multimodal classifier [2]. Since MM1 measures improvement over a baseline score, the measure is not sensitive to the specific type of classification accuracy measure being used, and hence, offers a flexible and more disciplined way of comparing classifier performance across models.

Table 3 presents the list of best unimodal and multimodal classifiers together with the associated MM1 effects. The estimates of MM1 scores show a number of interesting patterns. First, the average MM1 scores across the seven emotion classes is 20.2 for the cross-validated model, implying that the multimodal classifiers outperformed unimodal classifiers across all emotions. The highest improvements for multimodal classifiers were observed for the disgust and fear emotions.

The best performing unimodal classifier for neutral, happiness and sadness was the visual-only (V) classifier, while the best performing unimodal classifier for the anger, disgust, fear, and surprise was the audio-only (A) classifier. Among the multimodal classifiers, the trimodal ($V+A+T$) classifier outperformed bimodal classifiers across all emotions. The best performing bimodal classifier across all emotion classes was the $V+A$ classifier.

Given that disgust, fear, and surprise have high class imbalance, for rest of the analyses, we focus our attention on the emotion categories of neutral, happiness, sadness, and anger. We tested the robustness of the multimodal features using a random forest for these four emotion categories. The results from this analysis, detailed in Tables B1 and B2 in *Appendix B, available in the online supplemental material*, exhibit smaller effect sizes but are consistent with our key findings discussed here.

5.2 Contextual Effects: Gender and Duration

To enhance our understanding of the how gender of the speaker influences the inference of different emotions using a multimodal system, we performed an analysis of two gender categories (i.e., male vs. female). Using two human annotators, the dataset (2176 videos) was gender-annotated into 738 male-only and 1004 female-only videos. Around 434 videos could not be annotated with a specific gender as the gender of the speaker was either not discernible from the video even after inter-rater consultations, or was not uniquely identifiable on account of multiple speakers. For the successful gender annotations, the interrater agreement was 100 percent. Table B3 in *Appendix B, available in the online supplemental material*, provides the sample distribution across gender and emotion categories.

Table 4 below highlights the MM1 scores for the male-only and female-only subgroups extracted from the data.

Our results show that while multimodal classifiers offered significant[11] improvements over unimodal classifiers for both gender groups, there exists significant variance in the strength of this improvement, as discussed in detail in the next section. As mentioned earlier, we focus our attention only on the neutral, happiness, sadness, and anger emotions for this analysis, as the extent of class imbalance in these categories can be reasonably addressed using class weights and other techniques, e.g., SMOTE analysis [85]. Similar to the overall data, we tested the robustness of these findings for our gender subgroups using a random forest classifier. These additional results are presented in Tables B4 and B5 in *Appendix B, available in the online supplemental material*, and are consistent with our key findings discussed next.

The duration of an episode when an emotion is expressed can vary widely. As described earlier, the duration of episode might be longer or shorter depending on i) the modality e.g., emotions expressed through text take longer to express than through visuals, ii) the nature of medium e.g., short tweets and Vine videos vs. longer YouTube videos, and iii) the specific type of emotions, e.g., happiness takes shorter time to fully express than disgust.

In order to test the accuracy of the classifiers for the various emotion classes across short and long duration videos, we performed a quartile split on the training set based on the duration of videos, and categorized the first and fourth duration quartiles as short and long duration videos respectively. The first quartile comprised a total of 544 videos ranging in duration from 0.6 seconds to 4.4 seconds, while the fourth quartile comprised a total of 543 videos ranging in duration from 10.5 seconds to 30.7 seconds. Table B3 in the *Appendix B, available in the online supplemental material*, illustrates the sample distribution across duration and emotion categories.

We then trained an SVM classifier, similar to the one in the previous section, on these two groups of videos. We report the resulting MM1 scores for the two subgroups in Table 5 below. Similar to the gender subgroups, we tested and confirmed the robustness of this analysis using a random forest classifier; the additional results are presented in Tables B6 and B7 as part of *Appendix B, available in the online supplemental material*.

This analysis shows that multimodal classifiers outperformed unimodal classifiers across emotions for both shorter duration as well as longer duration videos. The only exception to this was the neutral class, where the best multimodal classifier trained on shorter duration videos performed at the same level as the best unimodal classifier (i.e.,

---

11. The difference in AUC scores between the best performing multimodal and unimodal classifier was statistically significant at 95% confidence level on a Wilcoxon signed-rank test.

TABLE 4
MM1 Scores for Male vs. Female Speakers Based on 10-Fold Cross-Validated AUC Scores

Modality	MM1 Score (%)			
	Neutral	Happiness	Sadness	Anger
Male Speakers	6.2	6.6	10.4	12.7
Overall	18.0	11.7	18.5	20.3
Best MM	$V+A+T$	$V+A+T$	$V+A+T$	$V+A+T$
Best UM	A	V	V	V
Female Speakers	1.7	4.8	5.8	11.1
Overall	18.0	11.7	18.5	20.3
Best MM	$V+A+T$	$V+A+T$	$V+A+T$	$V+A+T$
Best UM	V	V	V	V

MM1 = 0). For neutral and happiness, classifiers trained on longer duration videos reported higher *MM1* than those trained on shorter videos. However, for sadness and anger, classifiers trained on shorter duration videos reported a higher MM1 than those trained on longer videos.

6 DISCUSSIONS

6.1 Key Findings

Multimodal emotion features perform significantly better than unimodal features, but the modality effects differ across emotion types. One of the key insights from our analyses was the difficulty in classifying the neutral class. For the cross-validated model, the accuracy of classifying the neutral emotion was the lowest for the best unimodal (V, AUC = 0.583), best bimodal ($V+A$, AUC = 0.660) and trimodal ($V+A+T$, AUC = 0.688) classifiers, as compared to all other emotions. Furthermore, the MM1 score ($MM1_{neutral}$ = 18.0) for the neutral emotion was lower than the average MM1 score across all emotions ($MM1_{avg}$ = 20.2). The same pattern is also consistent across all the gender (i.e., male vs. female) and duration (i.e., shorter vs. longer) subgroups, where the MM1 scores for neutral emotion was found to be smaller than the MM1 scores for other emotions. These findings imply that for classifying neutral states, multimodal classifiers do not offer substantial improvements over unimodal classifiers, as compared to other emotions. More importantly, this reduction in improvement was most salient for shorter

duration videos where the best multimodal classifier showed no improvement over the best unimodal classifier ($MM1_{neutral}$ = 0). We found converging evidence for this using a random forest classifier as well ($MM1_{neutral}$ = 3.3).

This is consistent with recent work in affect recognition which emphasizes the complexity in conceptualizing what constitutes a neutral class and the problems with decoupling neutral-only videos from neutral segments of all other videos [84], [86]. Barros *et al.* reported that the neutral class exhibited the highest number of misclassifications in their analysis [86]. Similarly, Soleymani *et al.* reported high variance in pupillary response for participants when watching neutral scenes in videos [84]. Some past work has also shown that it is plausible for classifiers to confuse neutral state with emotions like happiness and sadness [87], [88]. In our analyses too, we found that the neutral and happiness emotion classes were the most difficult to classify, and showed lower improvements with multiple modalities as compared to other emotions. For all other emotions, multimodal accuracies were significantly higher than the unimodal accuracies for the training set.

Interestingly, we also observed that multimodal classifiers tend to work substantially better than unimodal classifiers for negatively valenced emotions like anger ($MM1_{anger}$ = 20.3 > $MM1_{avg}$ = 20.2), disgust ($MM1_{disgust}$ = 29.6 > $MM1_{avg}$ = 20.2), and fear ($MM1_{fear}$ = 24.1 > $MM1_{avg}$ = 20.2), although there is a possibility that these findings are specific to our data context, particularly given the imbalanced class distribution in disgust and fear

TABLE 5
MM1 Scores for Shorter vs. Longer Duration Videos Based on 10-Fold Cross-Validation

Modality	MM1 Score (%)			
	Neutral	Happiness	Sadness	Anger
Shorter Duration	0.0	6.3	10.3	17.9
Overall	18.0	11.7	18.5	20.3
Best MM	$V+A+T, V+A, V+T$	$V+A+T$	$V+A+T$	$V+A+T$
Best UM	V	V	V	A
Longer Duration	6.2	8.7	8.8	15.1
Overall	18.0	11.7	18.5	20.3
Best MM	$V+A+T$	$V+A+T$	$A+T$	$V+A+T$
Best UM	V	V	V	V

categories. However, we found that, for both SVM and random forest classifiers, negative emotions like sadness and anger exhibited better MM1 scores than positively valenced emotions like happiness. This can be partly because these emotions are complex or compound [26] and are unlikely to be fully learned through any one modality. In contrast, emotions like happiness have clear markers in at least one modality (e.g., facial smile) and can therefore be learned with reasonable accuracy through unimodal classifiers. This is consistent with recent work by [89] which found that facial expressions conveyed happiness the clearest, while vocal features conveyed anger better than other emotions. In our analyses too, we found that the most effective unimodal classifier for the happiness emotion was visual-only (V, AUC = 0.639).

Similar to fine-grained emotions, the improvement in multimodal performance was substantial in predicting the continuous dimensions like arousal and valence. This is consistent with recent studies that have shown high efficacy of multimodal systems in predicting valence and polarity [1], [3], [84]. In our analyses, we found that across emotions, bimodal and trimodal regression models outperformed unimodal regression models for both arousal and valence. Among the multimodal models, the trimodal model ($V+A+T$) outperformed all bimodal models. Further, we found that while audio-related features ($V+A$ and A) performed better than text-related features in predicting arousal, text-related features ($A+T$, $V+T$, T) performed better than audio features in predicting valence.

The gender of the speaker plays a moderating role on the multimodal features' performance. Our analyses on the OMG dataset also uncovered significant gender differences in emotion classification across various emotion classes. For instance, we found that multimodal classifiers outperformed unimodal classifiers for all emotions across the two genders (i.e., MM1 > 0). However, the average MM1 scores for male speakers ($MM1_{avg,male}$ = 9.0) outperformed those for female speakers ($MM1_{avg,female}$ = 5.9) across emotions. One possible explanation for this is that females, as young as 4 to 6 year old, have been found to express complex emotions [90], like sadness and anxiety, through a combination of subtle actions that are harder to detect, than for men who generally show limited reactions when expressing complex emotions. This is probably due to a mix of evolutionary limitations and restrictive social norms. For emotions like anger, where reactions like loud vocal expressions are relatively common for both men and women, the classifiers are less likely to show any difference. Recent studies looking at gender differences in emotion experience and perception have also found that women tend to be better than men at picking up certain emotions, and particularly negative emotions, as well as experiencing these emotions themselves [52], [53], [55]. However, our findings hint that, with the exception of anger, it might be challenging for multimodal classifiers to correctly classify other emotions expressed by female speakers.

Even though, in this study, we do not focus on gender effects in other negative emotions like fear and disgust due to the class imbalance in the available dataset, there have been recent studies that discuss relevant findings. For instance, and with disgust specifically, Al-Shawaf *et al*. provide a set of

evolutionary-functional reasons to explain why women consistently experience higher levels of disgust than men [54]. It is possible, therefore, that female speakers might be expressing this emotion more vividly than males, making it easier for classifiers to learn this difference. Similarly, for anger, an emotion closely related to disgust, our analyses show that unimodal and multimodal classifiers perform reasonably well for female speakers. Furthermore, for male speakers, the visual-only classifier was found to be the best performing unimodal classifier for happiness, sadness, and anger emotions, while the audio-only classifier performed better for neutral. For female speakers, however, the visual-only classifier was found to be better performing for all emotion classes.

Multimodal emotion features' effectiveness varies across episodic durations. The other contextual factor that we focused on was the duration of the emotion expression episode. We conjectured that the duration of a recorded emotion episode would be contingent on i) data modality, ii) nature of the medium, and iii) the type of emotion. While certain modalities, like text, require longer to fully onset a specific emotion (i.e., we might need to wait for a sentence to end to decipher the full meaning and intent from the speaker), other modalities like visual are quicker and more direct (e.g., a smile to convey happiness). Videos from certain online video sharing websites like YouTube, as is the case with the current dataset, tend to be longer in duration than videos from other platforms like Vine.

Our analyses on classifier performance for shorter vs. longer duration videos show that the average MM1 scores across emotions was comparable for the two duration groups ($MM1_{avg,shorter}$ = 8.6, $MM1_{avg,longer}$ = 9.7). However, we noticed variance across the specific emotion categories. For instance, and as mentioned earlier, multimodal classifiers failed to offer any improvement over unimodal classifiers for the neutral emotion when trained with *shorter* duration videos ($MM1_{neutral,shorter}$ = 0). For longer duration videos however, there was an improvement ($MM1_{neutral,longer}$ = 6.2). We also found that for neutral and positively valenced emotions, like happiness, the improvements in multimodal classification over unimodal classification were higher, as reflected by a higher MM1 score[12]. However, for negatively valenced emotions like sadness and anger, the opposite was true. This might partly be due to the nature of the specific emotions being expressed. For instance, while negative emotions like sadness and disgust might take longer to fully onset and peak, others such as happiness and surprise tend to be more immediate. More research is needed to fully understand the underlying mechanisms of why these duration effects manifest.

6.2 Implications, Limitations, and Future Research

Taken together, our analyses provide clear and converging empirical evidence supporting the multimodal approach to recognizing and analyzing emotions. More importantly, our findings highlight significant variances based on important contextual factors, like the gender of the speaker and the duration of the emotion episode.

---

12. The improvements in AUC scores were statistically significant at the 95% confidence level on a Wilcoxon signed–rank test.

Notably, our findings add to a growing stream of literature that cautions about the prevalence of biases in machine learning datasets and models [91], [92]. Our findings highlight that the relative performance of emotion recognition systems might vary substantially across genders, durations, emotions, and plausibly, the underlying feature extraction systems. Researchers and developers need to be aware of such variances especially if such systems are being used in a gender-sensitive context (e.g., employee selection based on video interviews).

There are a number of limitations in the present study that warrants future research. First, the current study uses text, speech and visual as the three data modalities, and uses all of these to compare the performance of the multimodal emotion recognition systems. We note that the feature combination technique used is platform-agnostic, i.e., while these modalities are common sources of emotion data, the relative importance and richness of these modalities might vary across platforms, e.g., Instagram content is likely to be richer in visual than text.

Second, and as pointed out in [48], the various modalities are not temporally aligned within the duration of the emotion episode. This might introduce errors in feature fusion, leading to higher misclassifications.

Third, while our study offers robust analytical insights based on real-world videos, the findings can be limited to the specific dataset we used. As we noted, the OMG dataset, though with a relatively data size and a good gender and duration diversity, included relatively fewer representations of disgust, fear and surprise which prevented us from extending our subgroup analyses to these classes. We also did not have access to the annotators' demographic distribution, which prevented us from studying the sensitivity of our findings, if any, to the annotators' gender.

Last but not the least, in this study, we selectively focused on two important contextual factors: gender and duration. Other factors like i) surrounding scenes, ii) physiological state (e.g., body temperature), iii) conversational contexts, etc., may also affect classifier performance. Similarly, the results are also plausibly sensitive to the performance and biases associated with specific tools, though reasonably effective, that we selected to generate the features. The relative performance of the classifiers might, therefore, also vary as a function of the emotion features extraction technologies, and inherit their innate biases. As technologies advance, future work in this direction shall further explore the effects of these factors on multimodal classifier accuracy.

Future research can look into building richer datasets, applying more advanced emotion analysis technologies, and studying abstractions that can explain the psychological underpinnings driving the associations between emotions, data modalities, genders and other contextual factors. A theoretically sound interpretation of multimodal emotion recognition can benefit not just the development of better systems, but also help preempt biases and explain several inconsistent findings reported in prior work.

7 CONCLUSION

This study leverages state-of-the-art visual, audio and text based emotion recognition systems, and presents an empirical analysis to further our understanding of the contextual factors affecting multimodal emotion recognition. While multimodal features consistently outperformed unimodal features, our findings revealed significant variance in improvement across different emotions. This study is also an early attempt at examining the effect of contextual factors like the gender of the speaker or the duration of the emotional episode, on classifier accuracy across modalities and emotion classes.

From an applied perspective, the relative strength of the unimodal or bimodal features in comparison to full, multimodal emotion recognition can provide a valuable reference to practitioners. In today's digital platforms, emotion cues are largely extracted from text-based unimodal (e.g., Tweets, Facebook comments), audio-text bimodal (e.g., call center recordings), or visual-text bimodal (e.g., Instagram stories and posts) content. Practitioners can derive a sense of confidence from our multimodal analysis to make an informed assessment of the relative predictive performance of the unimodal and bimodal data sources.

Emotions will continue to play a key role in shaping human experience and communication. The exploratory findings from this study offer valuable empirical insights to help understand and inform how future affective systems leveraging multimodal emotions can be improved with multiple data and communication modalities, different emotions of interest, as well as a wide range of contextual factors including but not limited to gender and duration of the expressive episode.

ACKNOWLEDGMENTS

REFERENCES

[1] S. K. D'Mello, N. Dowell, and A. Graesser, "Unimodal and multimodal human perception of naturalistic non-basic affective states during human-computer interactions," *IEEE Trans. Affect. Comput.*, vol. 4, no. 4, pp. 452–465, Fourth Quarter 2013.

[2] S. K. D'mello, and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM Comput. Surv.*, vol. 47, no. 3, pp. 1–36, 2015.

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

[4] B. Liu and S. S. Sundar, "Should machines express sympathy and empathy? Experiments with a health advice chatbot," *Cyberpsychol., Behav. Soc. Netw.*, vol. 21, no. 10, pp. 625–636, 2018.

[5] H. Cui, V. Mittal, and M. Datar, "Comparative experiments on sentiment classification for online product reviews," in *Proc. 21st Nat. Conf. Artif. Intell.*, 2006, pp. 1265–1270.

[6] M. Uljarevic and A. Hamilton, "Recognition of emotions in autism: A formal meta-analysis," *J. Autism Dev. Disord.*, vol. 43, no. 7, pp. 1517–1526, 2013.

[7] S. Alghowinem, R. Goecke, J. Epps, M. Wagner, and J. F. Cohn, "Cross-cultural depression recognition from vocal biomarkers," in *Proc. Interspeech*, 2016, pp. 1943–1947.

[8] D. McDuff, E. Kodra, R. el Kaliouby, and M. LaFrance, "A Large-scale analysis of sex differences in facial expressions," *PLoS One*, vol. 12, no. 4, 2017, Art. no. e0173942.

[9] N. H. Frijda, B. Mesquita, J. Sonnemans, and S. Van Goozen, "The duration of affective phenomena or emotions, sentiments and passions," *Int. Rev. Stud. Emot.*, vol. 1, pp. 187–225, 1991.

[10] K. Brans, and P. Verduyn, "Intensity and duration of negative emotions: Comparing the role of appraisals and regulation strategies," *PLoS One*, vol. 9, no. 3, 2014, Art. no. e92410.

[11] P. Verduyn and S. Lavrijsen, "Which emotions last longest and why: The role of event importance and rumination," *Motiv. Emot.*, vol. 39, no. 1, pp. 119–127, 2015.

[12] C. Darwin, and P. Prodger, *The Expression of the Emotions in Man and Animals*. New York, NY, USA: Oxford Univ. Press, 1998.

[13] E. M. Zemach, "What is emotion?," *Amer. Philos. Quart.*, vol. 38, no. 2, pp. 197–207, 2001.

[14] W. James, "What is an Emotion?," *Mind*, vol. 9, no. 34, pp. 188–205, 1884.

[15] G. L. Clore, K. Gasper, and E. Garvin, "Affect as information," in *Handbook of Affect and Social Cognition*. Mahwah NJ, USA: Lawrence Erlbaum Associates, 2001, pp. 121–144.

[16] S. P. Robbins and T. A. Judge, "Emotions and moods," *Organ. Behav.*, Pearson Education, 17th ed., pp. 138–171, 2017.

[17] J. M. Jenkins, K. Oatley, and N. Stein, *Human Emotions: A Reader*. Hoboken, NJ, USA: Wiley, 2000.

[18] N. H. Frijda, M. Lewis, and J. M. Haviland, "Moods, emotion episodes, and emotions," in *Handbook of Emotions*. New York, NY, USA: Guilford Press, 1993, pp. 381–403.

[19] W. James, F. Burkhardt, F. Bowers, and I. K. Skrupskelis, *The Principles of Psychology*. London, U.K.: Macmillan, 1890.

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

[21] J. Kagan, *What is Emotion? History, Measures, and Meanings*. New Haven, CT, USA: Yale Univ. Press, 2007.

[22] M. Rashid, S. A. R. Abu-Bakar, and M. Mokji, "Human emotion recognition from videos using spatio-temporal and audio features," *Vis. Comput.*, vol. 29, no. 12, pp. 1269–1275, 2013.

[23] G. Krell et al., "Fusion of fragmentary classifier decisions for affective state recognition," in *Proc. IAPR Workshop Multimodal Pattern Recognit. Social Signals Hum.-Comput. Interact.*, 2012, pp. 116–130.

[24] S. K. D'mello, and A. Graesser, "Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features," *User Model. User-Adapt. Interact.*, vol. 20, no. 2, pp. 147–187, 2010.

[25] P. Ekman, "An argument for basic emotions," *Cogn. Emot.*, vol. 6, no. 3–4, pp. 169–200, 1992.

[26] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[27] A. Ortony and T. J. Turner, "What's basic about basic emotions?," *Psychol. Rev.*, vol. 97, no. 3, pp. 315–331, 1990.

[28] T. J. Turner and A. Ortony, "Basic emotions: Can conflicting criteria converge?," *Psychol. Rev.*, vol. 99, no. 3, pp. 566–571, 1992.

[29] S. Wang, Y. Zhu, G. Wu, and Q. Ji, "Hybrid video emotional tagging using users' EEG and video content," *Multimedia Tools Appl.*, vol. 72, no. 2, pp. 1257–1283, 2014.

[30] T. Baltrušaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.

[31] M. S. Hussain, H. Monkaresi, and R. A. Calvo, "Combining classifiers in multimodal affect detection," in *Proc. 10th Australas. Data Mining Conf.*, 2012, pp. 103–108.

[32] K. Lu and Y. Jia, "Audio-visual emotion recognition with boosted coupled HMM," in *Proc. 21st Int. Conf. Pattern Recognit.*, 2012, pp. 1148–1151.

[33] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer, "Toward a minimal representation of affective gestures," *IEEE Trans. Affect. Comput.*, vol. 2, no. 2, pp. 106–118, Second Quarter 2011.

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

[35] J. Sonnemans and N. H. Frijda, "The structure of subjective emotional intensity," *Cogn. Emot.*, vol. 8, no. 4, pp. 329–350, 1994.

[36] N. H. Frijda et al., *The Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1986.

[37] R. K. Gupta, and Y. Yang, "Crystalfeel at SemEval-2018 task 1: Understanding and detecting emotion intensity using affective lexicons," in *Proc. 12th Int. Workshop Semantic Eval.*, 2018, pp. 256–263.

[38] M. Pantic, and L. J. Rothkrantz, "Toward an affect-sensitive multi-modal human-computer interaction," *Proc. IEEE*, vol. 91, no. 9, pp. 1370–1390, Sep. 2003.

[39] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 184–198, Second Quarter 2012.

[40] J. C. Lin, C. H. Wu, and W. L. Wei, "Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 142–156, Feb. 2012.

[41] B. Schuller, "Recognizing affect from linguistic information in 3D continuous space," *IEEE Trans. Affect. Comput.*, vol. 2, no. 4, pp. 192–205, Fourth Quarter 2011.

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

[43] O. AlZoubi, S. K. D'Mello, and R. A. Calvo, "Detecting naturalistic expressions of nonbasic affect using physiological signals," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 298–310, Third Quarter 2012.

[44] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *J. Multimodal User Interfaces*, vol. 3, no. 1–2, pp. 33–48, 2010.

[45] M. Wollmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image Vis. Comput.*, vol. 31, no. 2, pp. 153–163, 2013.

[46] M. Glodek et al., "Multiple classifier systems for the classification of audio-visual emotional states," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2011, pp. 359–368.

[47] B. de Gelder, H. K. M. Meeren, R. Righart, J. den Stock, W. A. C. de Riet, and M. Tamietto, "Beyond the face: Exploring rapid influences of context on face processing," *Prog. Brain Res.*, vol. 155, pp. 37–48, 2006.

[48] F. Lingenfelser, J. Wagner, J. Deng, R. Brueckner, B. Schuller, and E. André, "Asynchronous and event-based fusion systems for affect recognition on naturalistic data in comparison to conventional approaches," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 410–423, Fourth Quarter 2018.

[49] J. C. Kim and M. A. Clements, "Multimodal affect classification at various temporal lengths," *IEEE Trans. Affect. Comput.*, vol. 6, no. 4, pp. 371–384, Fourth Quarter 2015.

[50] H. Aviezer et al., "Angry, disgusted, or afraid? Studies on the malleability of emotion perception," *Psychol. Sci.*, vol. 19, no. 7, pp. 724–732, 2008.

[51] J. Brebner, "Gender and emotions," *Pers. Individ. Dif.*, vol. 34, no. 3, pp. 387–394, 2003.

[52] E. Hampson, S. M. van Anders, and L. I. Mullin, "A female advantage in the recognition of emotional facial expressions: Test of an evolutionary hypothesis," *Evol. Hum. Behav.*, vol. 27, no. 6, pp. 401–416, 2006.

[53] E. B. McClure, "A meta-analytic review of sex differences in facial expression processing and their development in infants, children, and adolescents," *Psychol. Bull.*, vol. 126, pp. 424–453, 2000.

[54] L. Al-Shawaf, D. M. G. Lewis, and D. M. Buss, "Sex differences in disgust: Why are women more easily disgusted than men?," *Emot. Rev.*, vol. 10, no. 2, pp. 149–160, 2018.

[55] W. A. Babchuk, R. B. Hames, and R. A. Thompson, "Sex differences in the recognition of infant facial expressions of emotion: The primary caretaker hypothesis," *Ethol. Sociobiol.*, vol. 6, pp. 89–101, 1985.

[56] D. Keltner and J. Haidt, "Social functions of emotions at four levels of analysis," *Cogn. Emot.*, vol. 13, no. 5, pp. 505–521, 1999.

[57] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The OMG-emotion behavior dataset," in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1–7.

[58] V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler, "Probabilistic approach to people-centric photo selection and sequencing," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2609–2624, Nov. 2017.

[59] S. Peng, L. Zhang, S. Winkler, and M. Winslett, "Give me one portrait image, I will tell you your emotion and personality," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 1226–1227.

[60] V. Vonikakis and S. Winkler, "Emotion-based sequence of family photos," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 1371–1372.

[61] V. Vonikakis, Y. Yazici, V. D. Nguyen, and S. Winkler, "Group happiness assessment using geometric features and dataset balancing," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 479–486.

[62] H. W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 443–449.

[63] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, First Quarter 2017.

[64] D.-Y. Huang et al., "Visual speech emotion conversion using deep learning for 3D talking head," in *Proc. Joint Workshop 4th Workshop Affect. Social Multimedia Comput First Multi-Modal Affect. Comput Large-Scale Multimedia Data*, 2018, pp. 7–13.

[65] D. Y. Huang et al., "Multimodal prediction of affective dimensions fusing multiple regression techniques," in *Proc. Interspeech*, 2017, pp. 162–165.

[66] X. Ouyang et al., "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 577–582.

[67] W. Ding et al., "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proc. 18th ACM Int. Conf. Multimodal Interact.*, 2016, pp. 506–513.

[68] W. Y. Quck, D. Y. Huang, W. Lin, H. Li, and M. Dong, "Mobile acoustic emotion recognition," in *Proc. IEEE Region 10th Conf .*, 2016, pp. 170–174.

[69] K. Poon-Feng, D.-Y. Huang, M. Dong, and H. Li, "Acoustic emotion recognition based on fusion of multiple feature-dependent deep Boltzmann machines," in *Proc. 9th Int. Symp. Chin. Spoken Lang. Process.*, 2014, pp. 584–588.

[70] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "Avec 2011–The first international audio/visual emotion challenge," in *Proc. Int. Conf. Affect. Comput. Intell. Interact.*, 2011, pp. 415–424.

[71] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychol. Sci.*, vol. 18, no. 12, pp. 1050–1057, 2007.

[72] J. Dennis and T. H. Dat, "Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2R'S system description for the ASpIRE challenge," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2015, pp. 518–524.

[73] V. H. Do, N. F. Chen, B. P. Lim, and M. A. Hasegawa-Johnson, "Multi-task learning using mismatched transcription for under-resourced speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 3, pp. 501–514, Mar. 2018.

[74] R. Tong, N. F. Chen, and B. Ma, "Multi-task learning for mispronunciation detection on Singapore children's Mandarin speech," in *Proc. Interspeech*, 2017, pp. 2193–2197.

[75] R. Duan and N. F. Chen, "Unsupervised feature adaptation using adversarial multi-task training for automatic evaluation of children's speech," in *Proc. Interspeech*, 2020, pp. 3037–3041.

[76] R. Duan and N. F. Chen, "Senone-aware adversarial multi-task training for unsupervised child to adult speech adaptation," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, Jun. 2021.

[77] N. F. Chen et al., "Strategies for Vietnamese keyword search," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2014, pp. 4121–4125.

[78] M. Nguyen, G. H. Ngo, and N. Chen, "Multimodal neural pronunciation modeling for spoken languages with logographic origin," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 2916–2922.

[79] M. Nguyen, G. H. Ngo, and N. Chen, "Hierarchical character embeddings: Learning phonological and semantic representations in languages of logographic origin using recursive neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 461–473, 2019.

[80] C. Ni, C.-C. Leung, L. Wang, N. F. Chen, and B. Ma, "Unsupervised data selection and word-morph mixed language model for Tamil low-resource keyword search," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process .*, 2015, pp. 4714–4718.

[81] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in tweets," in *Proc. 6th Joint Conf. Lexical Comput. Semantics*, 2017, pp. 65–77.

[82] R. K. Gupta, P. Bhattacharya, and Y. Yang, "What constitutes happiness? Predicting and characterizing the ingredients of happiness using emotion intensity analysis," in *Proc. AAAI Workshop Affect. Content Anal.*, 2019, pp. 152–163. [Online]. Available: https://dblp.org/rec/conf/aaai/GuptaBY19.html?view=bibtex

[83] R. K. Gupta and Y. Yang, "Predicting and understanding news social popularity with emotional salience features," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 139–147.

[84] M. Soleymani, M. Pantic, and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Second Quarter 2012.

[85] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[86] P. Barros, D. Jirak, C. Weber, and S. Wermter, "Multimodal emotional state recognition using sequence-dependent deep hierarchical features," *Neural Netw.*, vol. 72, pp. 140–151, 2015.

[87] C. Busso et al., "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proc. 6th Int. Conf. Multimodal Interfaces*, 2004, pp. 205–211.

[88] C. M. Lee et al., "Emotion recognition based on phoneme classes," in *Proc. 8th Int. Conf. Spoken Lang. Process.*, 2004, pp. 889–892.

[89] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Fourth Quarter 2014.

[90] L. R. Brody and J. A. Hall, "Gender and emotion in context," in *Handbook of Emotions*, New York, NY, USA: Guilford Press, 2008, vol. 3, pp. 395–408.

[91] J. P. Bajorek, "Voice recognition still has significant race and gender biases," *Harv. Bus. Rev.*, 2019. [Online]. Available: https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

[92] M. R. Costa-jussà, "An analysis of gender bias studies in natural language processing," *Nat. Mach. Intell.*, vol. 1, no. 11, pp. 495–496, 2019.

Prasanta Bhattacharya received the PhD degree in information systems from the National University of Singapore. He is currently a scientist and innovation lead with Social and Cognitive Computing Department, Institute of High Performance Computing, A*STAR, Singapore. He currently serves as an adjunct assistant professor with the Department of Analytics and Operations, NUS Business School. His research interests include social network analyses and behavioral analytics.

Raj Kumar Gupta received the PhD degree in computer science from Nanyang Technological University, Singapore. He is currently a senior scientist with Social and Cognitive Computing Department, Institute of High Performance Computing, A*STAR, Singapore. He has more than five years industry experience on software development, visual analytics, and embedded systems. His research interests include natural language processing, sentiment and emotion analysis, and multimodal analysis.

Yinping Yang received the PhD degree in information systems from the National University of Singapore. She is currently a senior scientist and the principal investigator with A*STAR Digital Emotions programme, Institute of High Performance Computing, A*STAR, Singapore. She manages industry projects and advises tech start-ups in emotion analytics and e-negotiation technologies. Her research interests include intelligent negotiation systems, emotion recognition and sentiment analysis, and strategic foresight.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.