EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children

Behnaz Nojavanasghari
Computer Science
University of Central Florida
Orlando, FL, USA
behnaz@eecs.ucf.edu

Tadas Baltrušaitis
LTI, Carnegie Mellon University
Pittsburgh, PA, USA
tbaltrus@cs.cmu.edu

Charles E. Hughes
Computer Science
University of Central Florida
Orlando, FL, USA
ceh@eecs.ucf.edu

Louis-Philippe Morency
LTI, Carnegie Mellon University
Pittsburgh, PA, USA
morency@cs.cmu.edu

ABSTRACT

Automatic emotion recognition plays a central role in the technologies underlying social robots, affect-sensitive human computer interaction design and affect-aware tutors. Although there has been a considerable amount of research on automatic emotion recognition in adults, emotion recognition in children has been understudied. This problem is more challenging as children tend to fidget and move around more than adults, leading to more self-occlusions and non-frontal head poses. Also, the lack of publicly available datasets for children with annotated emotion labels leads most researchers to focus on adults. In this paper, we introduce a newly collected multimodal emotion dataset of children between the ages of four and fourteen years old. The dataset contains 1102 audio-visual clips annotated for 17 different emotional states: six basic emotions, neutral, valence and nine complex emotions including curiosity, uncertainty and frustration. Our experiments compare unimodal and multimodal emotion recognition baseline models to enable future research on this topic. Finally, we present a detailed analysis of the most indicative behavioral cues for emotion recognition in children.

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI)**; •**Applied computing** → *Psychology; Education;*

Keywords

Emotion Recognition, Nonverbal Behavior Analysis, Audio-Visual Sensing, Facial Analysis

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2993168

Figure 1: Example frames of EmoReact. This figure shows wide variations of facial expressions in EmoReact.

1. INTRODUCTION

The recognition of emotions helps humans with their development of social skills and successful communication and it play a significant role in perception and decision making in everyday life [14, 33, 19]. With newly emerging domains such as emotionally intelligent robots [37], affect-sensitive human computer interaction design [41], affect-aware tutors [15], automatic emotion recognition is becoming a part of human life. With rapid development of web technologies, distance learning is becoming widely used by students. Most of these users are young adults, teenagers and children [13] who benefit from online classes for completing their education. Automatic emotion recognition using facial expressions [24], gesture and posture [28], speech [44] and text mining [10, 17] has received great attention from computing communities in recent years.

Although there has been a considerable amount of research on emotion recognition of adults, emotion recognition in children has been understudied. Focusing on emotion recognition for children will enable us to build child friendly systems that enhance the quality of distant education and

Table 1: Comparison of EmoReact with other publicly available datasets of children. EmoReact is the largest dataset of its kind both in size of data and number of annotated emotion labels. It also contains a wide age range of children from both genders. A - audio, V - video, P - physiological, I - images F- females.

Properties	EmoReact	MMDB[43]	CAFE[39]	Radboud[36]	NIMH[25]	Dartmouth[16]
Modalities	A/V	A/V/P	I	I	I	I
# of Samples	1102 V	160 V	1192 I	80 I	482 I	640 I
# of Children	63	121	154	10	59	80
# of Labels	17	2	7	8	5	8
Gender	51 % F	-	58 % F	60 % F	63 % F	50 % F
Age Range	4-14	1-2	2-8	8-12	10-17	6-16

care, providing a significant long-term return on investment. When students work in cooperative groups, they develop an understanding of the collective purpose and the need to constructively solve problems while supporting each other's learning [32]. In such context, automatic emotion recognition can be helpful to understand learning in children.

Recognizing emotions in children can be very challenging due to a number of reasons: 1) Rapid motions: Children move more compared to adults, which makes it challenging to successfully track their face and body gestures. 2) Occlusion: There is partial or full occlusion of faces caused by facial orientation and objects or body parts such as hair bangs, which are very common for children. 3) Lack of resources: There are few publicly available datasets of children; only one contains videos and all other ones contain only images.

In this work we present a newly collected multimodal emotion dataset of children between the ages of four and fourteen years old that contains 1102 videos; the biggest dataset of its kind. These videos are annotated for 17 affective states, including six basic emotions (happiness, sadness, surprise, fear, disgust, and anger), neutral, valence and nine complex emotions including curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment and frustration. We provide a detailed analysis of the visual and vocal behaviors shown by children expressing these emotions. Finally, we establish unimodal and multimodal baselines for classifying the emotion labels and compare the performance of classifiers across emotions. In Figure 1 you can see example frames from the dataset showing the expression of different emotions. This dataset will be available to the public.

2. RELATED WORK

Automatic emotion recognition has been a topic of interest for a long time due to its wide applications. There has been a great amount of research on emotion recognition using visual signals [53], acoustic signals [44], physiological signals [34] and verbal cues [51]. However, most of these works focus on emotion recognition for adults.

One of the primary reasons is the lack of publicly available datasets for emotion recognition in children (see Table 1). Of the existing datasets, only the MMDB dataset contains video recordings of young children between one and two years old, with annotations for responsiveness and engagement. Although this dataset contains annotated affective states as well, its goal is to study the social behavior of young children. Also, as the subjects are toddlers, this makes it hard to analyze the vocal channel due to their very limited vocabularies. The findings on the MMDB dataset might not generalize well to older children who have complex vocabularies and will rely partially on those words to convey their emotions.

A number of other datasets contain only still images of children. The Child Affective Facial Expression (CAFE) [39] dataset contains images from children between two and eight years old. This dataset includes labels for six basic emotions and neutral. The NIMH dataset [25] has 482 photos of four basic emotions (fearful, angry, happy, and sad) and neutral of children between ten and seventeen years old with both directed gaze and averted gaze. The Dartmouth dataset [16] contains 640 posed images of children between six and sixteen years old, which contains six basic emotions, neutral and afraid emotion labels. This dataset also has five different images of the same emotion per subject, which were taken from different angles. A part of Radboud faces dataset [36] also contains images of children. This dataset contains 80 images of children between eight and twelve years old posing six basic emotions, neutral and contempt and three images were taken from each child per emotion from frontal, left and right angles. Although there are emotions that can be conveyed through a single image, human behavior is dynamic and there has been research that shows the study of behavior in time versus a single image can lead to different conclusions [30]. Also, when collecting some of these datasets, children have been asked to pose the desired emotion, and so the expression of emotions might not be natural and lack variability.

To the best of our knowledge, EmoReact is the first multimodal dataset that covers a wide age range of children annotated for 17 affective states. Table 1 shows a comparison between EmoReact and other publicly available datasets of children.

3. EMOREACT DATASET

In this section we will describe how we collected the dataset. Then we will explain our annotation scheme for providing the labels. Finally, we will provide some statistics and analysis of the dataset and its associated annotations.

3.1 Dataset Collection

Online social media websites such as YouTube and Facebook have become unbounded sources of multimedia content. YouTube has become a significant source of video data where hundreds of hours of new videos are uploaded every minute[1] . We have selected React channel[2] from YouTube as the source from which we downloaded videos of children who are reacting to different subjects. These videos contain

[1]https://en.wikipedia.org/wiki/YouTube(accessed 05/2016)

[2]https://www.youtube.com/user/React(accessed 05/2016)

Table 2: Number of videos containing each emotion and number of people who have expressed that emotion in EmoReact.

Emotion Labels	Number of Videos	Number of Children
Curiosity	385	51
Uncertainty	344	53
Excitement	355	49
Happiness	604	60
Surprise	298	49
Disgust	137	35
Fear	50	20
Frustration	131	31

children between the ages of four to fourteen years old, from different races and both genders. We found this source to be a very rich resource to study emotional expressions in children and we have downloaded videos of children reacting to 37 subjects that include food, technology, YouTube videos and gaming devices. Through these videos, children performed the following five tasks: (1) getting to know the subject by its being shown, (2) being asked a question about the subject, (3) answering a question about the subject, (4) being told a fact about the subject and reacting to it, and (5) explaining one's opinion about the subject. Each of the initially downloaded videos included multiple children reacting to a subject. We have manually segmented the videos into short clips, with an average length of five seconds, using ELAN[6], so that each video clip contains only one child who is reacting to one specific subject. From this segmentation step, we kept only the videos that are longer than three seconds resulting in a total of 1254 clips.

In order to enable both person-independent and person-specific analysis, we have annotated the identity of each child in each of the 1254 video clips. We used the approach proposed by Florian et al [47], followed by a manual inspection and correction. This resulted in a dataset of 63 children from which 32 are female and 31 are male.

3.2 Emotion Annotations

In recent years, researchers from psychology, education and computer science have turned their attention to the role of affect in education and learning. Our choice of emotion labels is done with an emphasis on affective states that are important for learning and education based on previous research [9]. The full list of our labels includes six basic emotions (anger, disgust, fear, happiness, sadness and surprise), neutral, curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, frustration and valence. We requested annotators to also annotate the gender and judge the naturalness of the reactions.

To obtain our labels, we recruited crowd workers from the online crowd sourcing platform, Amazon's Mechanical Turk (MTurk) [7]. Selection of the workers was performed through a pre-study. For the pre-study we hired nine different workers from people who signed up to complete the study based on their experience. We divided the workers into three groups and assigned 20 videos to each group. The pre-study had all of the questions regarding emotion labels that were in the final study and several questions such as identifying the gender of the children to assess the quality of the annotations.

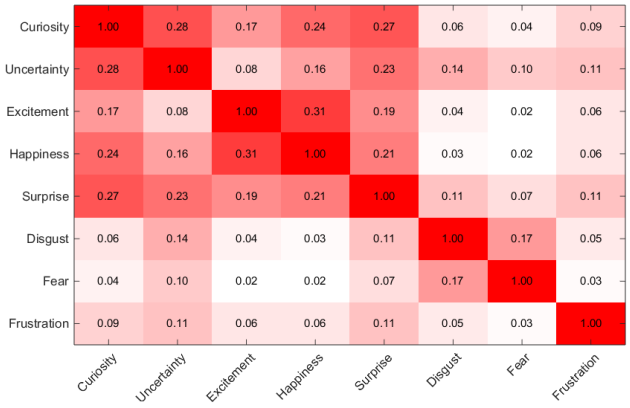

Figure 2: Co-occurrence between different affective states.

After analyzing the agreement level between workers in each group of the pre-study, we hired two groups out of three initial groups for our main study, selecting those who reached the highest inter-coder agreement. The final set of six workers included three female and three male workers to reduce gender bias [21]. Each video was annotated by three independent workers for seventeen labels. The interface for annotations contained the definitions of each label for consistency and, as a test of the rater's vigilance and rational decision making, we included a question about the gender of the child in the video for pruning random annotations. All emotions except valence are annotated on a 1-4 Likert scale where 1 shows the absence of emotion and 4 shows the intense presence of the emotion(with 2 and 3 showing little and moderate expression of the emotion). Valence is annotated on a scale of 1-7 where 1 shows the most negative emotion and 7 shows the most positive emotion.

After collecting the labels from MTurk, we used Kripendorff's alpha [35] to evaluate the agreement level between workers. In this step, we removed 152 videos, where it looked like the annotators were not vigilant. This processing led us to our final set of 1102 videos. The agreement levels for different labels are as follows: Happiness: 0.57; Surprise: 0.63; Disgust: 0.61; Fear: 0.43; Curiosity: 0.41; Uncertainty: 0.47; Excitement: 0.43; Frustration: 0.54, Exploration: 0.24; Confusion: 0.29; Anxiety: 0.31; Attentiveness: -0.16; Anger: 0.28; Sadness: 0.23; Embarrassment: 0.09; Valence: 0.65; Neutral: 0.37. Agreement level between 0.4 and 0.6 shows moderate agreement and values between 0.6 and 0.8 show substantial agreement level between raters [29]. These agreement levels compare favorably to previous work in affective computing [20].

We decided to use the eight emotion labels that reached the highest levels in coder agreement (greater than 0.4) for initial analysis of the dataset and our experiments in this paper.

3.3 Dataset and Annotations Statistics

In Table 2 you can see information regarding the number of videos for each emotion label in EmoReact as well as the number of different children who have expressed each emotion in the dataset. Each emotion is expressed by a large number of people, which will enable researchers to build computational models that generalize well across unseen in-

dividuals. Most of the previous work on emotion recognition, has assumed an exclusive label for each video/image. However, we allow each video to have more than one emotion label. Also, we have analyzed the co-occurrence of emotions and have provided the co-occurrence ratio of emotions in Figure 2. As an example, it is interesting to note that curiosity has mostly appeared with uncertainty, surprise and fear. Curiosity has been defined as a need, thirst or desire for knowledge about something [2], which can mean the curious person is uncertain about some aspects about a topic. Discovering knowledge about that topic can be surprising, especially if it is contradictory to one's previous beliefs, and can cause happiness. This finding is also consistent with some of the findings of previous research [3, 27, 45, 49] about the co-occurrence of curiosity with other emotions.

4. MULTIMODAL BEHAVIOR ANALYSIS

Emotion can be expressed through both vocal and visual cues.There has been very little work analyzing the detailed behaviors happening during children's expression of emotional states[38]. Our goal is to better understand the visual and acoustic behaviors correlated with presence of specific emotional states. For the purpose of this analysis, presence of an emotion label is defined as at least two workers agreeing that it is present (above or equal to two on a four-point Likert scale).

We have performed a t-test between the video clip instances that the emotion was annotated as present and the instances where it was absent, to identify the most indicative behaviors. Please note that we have corrected the p-values with respect to number of our features because of multiple comparisons [5]. Also, we complement the p-values with the effect size measured with Hedges' g [23]. Reported effect sizes shows the effectiveness and practical significance of a particular behavior.

In this section, we first analyze the visual behaviors and then we provide analysis of acoustic behaviors and summarize some of our findings about the most predictive behaviors for each emotion.

4.1 Visual Behavior Analysis

As the face discloses important information during one's emotional responses, we decided to study the following types of visual behaviors:

- **Facial Action Units:** These are the most basic independent visible movement of the facial musculature, which are known to be a strong indicator of the emotion present in one's face [26]. The list of actions units we have used in this paper is as follows: Inner brow raiser, Outer brow raiser, Brow lowerer, Upper lid raiser, Cheek raiser, Lip lightener, Nose wrinkler, Upper lip raiser, Lip corner puller, Dimpler, Lip corner depress, Chin raiser, Lip stretcher, Lip tightener, Lips part, Jaw drop, Lip suck, Blink.

- **Head Position and orientation:** We have used head pose to get information about the direction of the face (toward the director, object or away). This set of features gives us information about the head orientation and movement in the x, y, and z axes which are important indicators for some behaviors such as head nods and head shakes, which have been shown to be indicators of agreement and disagreement [4].

- **Non Rigid Shape parameters:** These parameters are the result of applying PCA on facial landmarks as they control the movements of the facial points such as widening the eyes or opening mouth which can be important in expressing some emotions such as fear.

In Table 3 you can see the most predictive visual behaviors for each emotion label. All of these behaviors have a p-value smaller than 0.001, which is p-value corrected for a number of comparisons [5]. We have included the effect size as well to indicate the direction and size of the effect. Based on our analysis, head orientation and action units are the most differentiating behaviors for recognizing our emotion set.

For example, in the case of presence of curiosity, let us consider a child who is examining an object such as an iPad, a telephone or a movie. The overall head rotation is lower and the horizontal gaze shift is more among curious children in compared to non-curious children. This might be happening because the child is focused on the new object and is trying to get more information about it by looking at different parts of it. In the case of uncertain children, they move their head around more - as indicated by higher variance in head translation. This can be happening because children move their heads examining the object to discover new things about it and find an answer for their uncertainty.

It is important to note that in each video a child can show more than one emotion and some emotions co-occur more often. For example, disgust happens often when the subject of the video is food. If children are disgusted by the food, they can have different reactions depending on the type of the food. If they would like to try the food they could be happily disgusted; if they are afraid of the food, possibly because of its strange appearance, they could be fearfully disgusted or if they have never seen any similar food they can be disgustedly surprised.

There has been previous research on compound emotions [22] where researchers have defined emotions like disgustedly surprised, fearfully disgusted, happily disgusted, etc. In Table 3 you can see that one of the most predictive cues for disgust is the higher variance of vertical gaze shift, which shows looking away behavior, lip stretcher, which is known to happen mostly with fear [49], nose wrinkle, which usually occurs with disgust [40] and upper lid raiser, which is often associated with fear or surprise [12]. We believe that co-occurrence of emotions can cause overlap between indicators of different emotions that can be considered as a compound emotion (disgust and surprise, disgust and excitement).

4.2 Acoustic Behavior Analysis

Some emotional states may be best recognized with acoustic features, even if visual cues are provided and are not noisy. We studied some of the most successful acoustic features for emotion recognition based on previous research [50].

- **Voice quality features:** Normalized amplitude quotient (NAQ), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), quasi-open quotient (QOQ) difference between the first two harmonics (H1-H2), and peak-slope. These features measure the pressedness, tenseness, creakiness or breathiness of speech.

- **Prosody:** Pitch/Fundamental Frequency (F0) indicates the pitch information in one's speech.

Table 3: Summary of most predictive visual behaviors for each emotion label. Trend shows the direction of effect size for that behavior in presence of emotion label ($p^{***} < 0.001$).

Emotion	Feature	Stat	Hedges' g	Trend
Curiosity	HeadRot(rx)	μ	-0.54	$\downarrow$
	H Gaze shift	μ	0.48	$\uparrow$
Uncertainty	Scale	δ	-0.26	$\downarrow$
	HeadTrans(tx)	δ	0.22	$\uparrow$
Excitement	Lip corner puller	δ	0.42	$\uparrow$
	Lip Stretcher	δ	0.34	$\uparrow$
	Cheek raiser	δ	0.28	$\uparrow$
Happiness	HeadRot(rz)	μ	-0.29	$\downarrow$
	Upper lip raiser	δ	-0.22	$\downarrow$
Surprise	Lip corner puller	δ	-0.38	$\downarrow$
	Upper lip raiser	μ	-0.25	$\downarrow$
Disgust	V Gaze shift	δ	0.48	$\uparrow$
	HeadRot(rz)	δ	-0.43	$\downarrow$
	Lip stretcher	δ	0.35	$\uparrow$
	Nose wrinkler	μ	0.33	$\uparrow$
	Upper lid raiser	μ	0.33	$\uparrow$
Fear	Non-rigid P 22 *	μ	0.55	$\uparrow$
	Non-rigid P 8 *	δ	-0.54	$\downarrow$
Frustration	HeadRot(rx)	μ	-0.72	$\downarrow$
	Lip corner puller	μ	-0.42	$\downarrow$

* These parameters correspond to widening the eyes and opening the mouth respectively.

- **MFCC:** Mel-Frequency Cepstral Coefficient is widely used in the speech recognition community and has been shown to be successful in emotion recognition [46]. MFCC features mimic the behavior of human ears by applying cepstral analysis and measuring the perceived frequencies (pitch) of a pure tone to its actual measured frequency. We have used MFCC 1-24 in our experiments.

We have also considered the statistics of voiced segments (VUV) as a feature to capture length of speech in each clip. Table 4 shows our detailed analysis of acoustic behavior indicators. All the behaviors have p-values smaller than 0.001, which is a corrected p-value for acoustic behaviors based on number of comparisons [5].

The acoustic behaviors that have the most effect on our set of emotion labels are MFCC features, voice quality features such as H1H2, MDQ and prosody features such as F0. H1H2 and MDQ features correspond to the breathiness and tenseness of the speech signal and previous research [8, 15, 42] has shown their effectiveness in emotion recognition. Breathy voice was indicative of emotions such as uncertainty and sadness while more tense and creaky voice was indicative of emotions like fear and disgust [42]. In our analysis, we found that, in cases where fear and disgust are present, the mean of MDQ is lower and the mean of F0 is higher, which shows the tense and creaky voice and high voice energy (pitch) respectively.

5. EXPERIMENTAL METHODOLOGY

To assess the usefulness of EmoReact in building person independent automatic emotion recognition systems for children we performed a number of experiments using a set of machine learning approaches. Our experiments show the comparison between our models and basic approaches, importance of each modality in predicting each emotion label, and a comparison between unimodal and multimodal approaches.

Table 4: Summary of most predictive acoustic behaviors for each emotion label. Trend shows the direction of effect size for that behavior in presence of emotion label ($p^{***} < 0.001$).

Emotion	Feature	Stat	Hedges' g	Trend
Curiosity	MFCC 14	δ	-0.44	$\downarrow$
	H1H2	δ	-0.35	$\downarrow$
	QOQ	μ	0.31	$\uparrow$
Uncertainty	MFCC 14	δ	-0.35	$\downarrow$
	H1H2	δ	-0.28	$\downarrow$
Excitement	QOQ	δ	0.70	$\uparrow$
	F0	μ	0.50	$\uparrow$
Happiness	MFCC 3	μ	0.24	$\uparrow$
	PSP	δ	0.22	$\uparrow$
Surprise	MFCC 10	μ	0.48	$\uparrow$
	VUV	δ	0.40	$\uparrow$
	PSP	δ	-0.39	$\downarrow$
Disgust	MDQ	μ	-0.70	$\downarrow$
	MFCC 21	δ	-0.63	$\downarrow$
	F0	μ	0.60	$\uparrow$
Fear	F0	μ	0.81	$\uparrow$
	MDQ	μ	-0.69	$\downarrow$
Frustration	MFCC 15	δ	0.68	$\uparrow$
	F0	μ	0.40	$\uparrow$

5.1 Features

We used OpenFace [1], which is an open source tool, to extract visual features. We selected the valid frames where the faces are successfully detected and are close to frontal. In total 72.88% of the frames in EmoReact are successfully processed by OpenFace. To extract acoustic features we used a publicly available software tool - COVAREP[18]. Acoustic descriptors are computed on the voiced segments of the audio streams and every 10 millisecond. In total 76.52 % of the videos have valid speech signals.

After extracting the raw features from both modalities we computed the mean and standard deviation as a way of summarizing from short windows to an entire video.

5.2 Baseline Models

In order to show the prediction performance of conventional classifiers and provide baseline models for future research we have trained Naive Bayes, and linear and radial basis function kernel SVM classifiers using visual, acoustic, and audio-visual features.

5.3 Implementation Details

We separated EmoReact into three subsets: Training set, which contains 432 videos; validation set, which has 303 videos; and test set, which has 367 videos. These sets are defined in a person independent manner for better generalizability of our models and conclusions. Emotions in each set have been demonstrated by 21 different children and the distributions of emotions in all three sets are similar.

Table 5: Comparison between our classifiers and baseline models, reporting average performance across emotions.

Method	Audio		Visual		Audio-Visual	
	AUC ROC	F_1	AUC ROC	F_1	AUC ROC	F_1
Majority voting	0.50	- *	0.50	- *	0.50	- *
Random	0.50	0.50	0.50	0.50	0.50	0.50
Naive Bayes	0.57	0.55	0.59	0.57	0.61	0.61
Linear SVM	0.61	0.66	0.62	0.66	0.63	0.65
RBF SVM	0.61	0.69	0.62	0.68	0.64	0.69

* Please note that that skew normalized F1 is not defined for majority voting.

We used Naive Bayes, linear SVM and radial basis function kernel SVM as our baseline models [11]. We used the validation set for selecting the hyper parameters of SVM (i.e., $\gamma = 2^{-}5, ..., 2^{5}$ and $C = 10^{-}5, ..., 10^{5}$ parameters).

To address the challenge of imbalanced emotion labels we followed two approaches: (1) under-sampling negative examples, and (2) using an ensemble classifier. We randomly selected a subset of negative examples that was equal to the number of positive samples. We repeated this procedure 100 times to create multiple balanced training sets. A predictive model was trained on each of the 100 balanced training sets. The ensemble classifier was created by using a simple majority voting scheme to integrate the prediction of each individual prediction model. The same approach was followed for all baseline models presented in this paper.

For building our unimodal models we used only one set of features (visual or acoustic). For multimodal approaches we explored the following options:

Early fusion: We concatenated the visual and acoustic features and used the new feature set as input to classifier.

Late fusion: We took a majority vote between ensembles of visual and acoustic classifiers.

Hybrid fusion: We took a majority vote between ensembles of visual, acoustic and audio-visual classifiers.

6. RESULTS AND DISCUSSIONS

In this section, we introduce our error metrics, present the results of baseline models, show the performance of different modalities in predicting each emotion label, and compare unimodal and multimodal approaches. We also provide a discussion of our findings.

Metrics: Following prior work in facial action unit recognition [31], where imbalanced data is also an important issue, we used two error metrics robust to imbalanced datasets: Area under the curve of ROC, which shows area under the curve of the true positive rate as the function of the false positive rate and is the only metric that has been shown to be robust to imbalance in the dataset; and skew normalized F1 which can be interpreted as a weighted average of precision and recall values on a balanced test set.

6.1 Baseline Results

Table 5 summarizes the performance of baseline models on the EmoReact dataset. We have compared the performance of Naive Bayes, linear and RBF SVM classifiers with majority voting and a random guess classifier and have reported both area under ROC and F1. Consistently the baseline models are outperforming majority voting and random classifiers on both metrics, demonstrating the usefulness of EmoReact in building emotion recognition models for children. Both SVM models outperform Naive Bayes, with RBF SVM showing slightly better performance than Linear SVM.

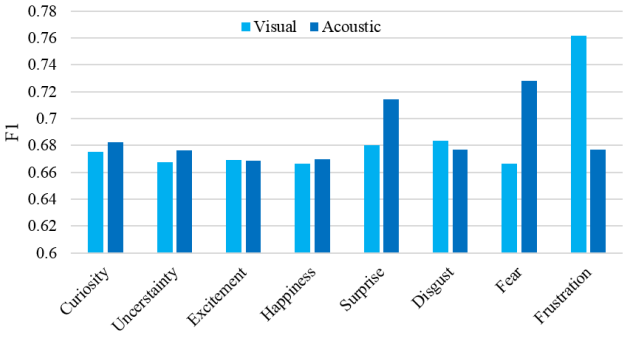

Figure 3: Comparison between visual and acoustic models in predicting emotions.

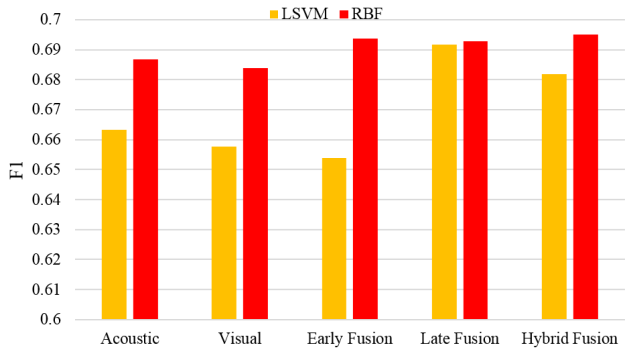

Figure 4: Comparison between unimodal and multimodal approaches reporting average performance across emotions.

6.2 Visual vs. Acoustic Modality

Figure 3 shows the performance of our classifiers based on each modality in predicting emotions. The purpose of this experiment is to show the performance of each modality in predicting emotions.

We trained classifiers using visual only and acoustic only features. As you can see in Figure 3, there are emotions such as fear and surprise, where children have communicated the emotions mostly using speech signals and acoustic behaviors have been more powerful for predicting the emotion labels. On the other hand, there are emotions such as frustration and excitement, where children have expressed them mostly by visual behaviors and visual models are doing a better job at predicting those emotions. Also, as previous research suggests [48, 52], the performance of visual classifiers might be affected by challenges such as open mouth, facial occlusions and rapid movements. It is interesting to see that, in the case of fear, children cover their mouths which causes partial occlusion or in the case of surprise, open mouth is a common case. Also, audio signals can be affected by unwanted sounds such as dropping an object while speaking or background music. These challenges may lead to noise in

face tracking, action unit recognition and acoustic features.

6.3 Unimodal vs. Multimodal

The aim of this experiment is to compare the performance of unimodal and multimodal approaches. Figure 4 shows the results of these comparisons. It can be seen that most of the multimodal approaches are performing better than the unimodal ones.

We believe that the reason for better performance is due to the fact that each modality is more powerful in predicting certain emotions. Also, one modality can compensate for another one, when the other modality is missing or noisy.

7. CONCLUSIONS AND FUTURE WORK

In the present work, we introduced a new multimodal emotion dataset of children that contains 1102 video clips, which are annotated for 17 emotion labels. We provided a detailed analysis of visual and acoustic behaviors that are the most indicative cues for the presence of emotions. Furthermore, we presented unimodal and multimodal approaches for predicting emotions, which will be the baseline for all future research on EmoReact. Our results suggest that each modality can be more successful in predicting certain emotions and modalities can work in a complementary fashion. Finally, we have shown that different multimodal approaches (early, late, hybrid fusion) have similar performances but are more successful for predicting emotions compared to unimodal models. As future work, we intend to add captions to EmoReact to make research on verbal modality possible as well. We are also interested to explore the gender effect on emotion expression in children.

8. ACKNOWLEDGMENTS

We would like to thank Dr. Sunghyun Park for his insightful advices during our crowd sourcing and annotation process.This material is based upon work partially supported by The Heinz Endowments (co-PI: Cassell, Hammer and Morency) and the Bill & Melinda Gates Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of The Heinz Endowments or the Bill & Melinda Gates Foundation, and no official endorsement should be inferred.

9. REFERENCES

[1] T. Baltrusaitis, P. Robinson, L.-P. Morency, et al. OpenFace: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016.

[2] D. E. Berlyne. Conflict, arousal, and curiosity. 1960.

[3] N. Bosch and S. D'Mello. It takes two: momentary co-occurrence of affective states during computerized learning. In *International Conference on Intelligent Tutoring Systems*. Springer, 2014.

[4] K. Bousmalis, L.-P. Morency, and M. Pantic. Modeling hidden dynamics of multimodal cues for spontaneous agreement and disagreement recognition. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011.

[5] G. E. Box, W. G. Hunter, J. S. Hunter, et al. Statistics for experimenters. 1978.

[6] H. Brugman, A. Russel, and X. Nijmegen. Annotating multi-media/multi-modal resources with elan. In *LREC*, 2004.

[7] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 2011.

[8] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004.

[9] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on affective computing*, 2010.

[10] E. Cambria, B. Schuller, Y. Xia, and C. Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 2013.

[11] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2011.

[12] A. V. Clark. *Psychology of moods*. Nova Publishers, 2005.

[13] J. T. Colorado and J. Eberle. Student demographics and success in online learning environments. *Emporia State Research Studies*, 2010.

[14] A. J. Cotugno. Social competence and social skills training and intervention for children with autism spectrum disorders. *Journal of autism and developmental disorders*, pages 1268–1277, 2009.

[15] A. Cullen, J. Kane, T. Drugman, and N. Harte. Creaky voice and the classification of affect. *Proceedings of WASSS, Grenoble, France*, 2013.

[16] K. A. Dalrymple, J. Gomez, and B. Duchaine. The dartmouth database of children's faces: acquisition and validation of a new face stimulus set. *PloS one*, 2013.

[17] T. Danisman and A. Alpkocak. Feeler: Emotion classification of text using vector space model. In *AISB 2008 Convention Communication, Interaction and Social Intelligence*, volume 1, 2008.

[18] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. COVAREP -A collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014.

[19] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

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

[21] U.-S. Donges, A. Kersting, and T. Suslow. Women's greater ability to perceive happy facial emotion automatically: gender differences in affective priming. *PLoS One*, 2012.

[22] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 2014.

[23] J. A. Durlak. How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology*, 2009.

[24] J. Edwards, H. J. Jackson, and P. E. Pattison. Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical psychology review*, 2002.

[25] H. L. Egger, D. S. Pine, E. Nelson, E. Leibenluft, M. Ernst, K. E. Towbin, and A. Angold. The nimh child emotional faces picture set (nimh-chefs): a new set of children's facial emotion stimuli. *International Journal of Methods in Psychiatric Research*, 2011.

[26] P. Ekman. Strong evidence for universals in facial expressions: a reply to russell's mistaken critique. 1994.

[27] M. W. Gallagher and S. J. Lopez. Curiosity and well-being. *The journal of positive psychology*, 2007.

[28] H. Gunes and M. Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 2007.

[29] K. A. Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 2012.

[30] M. Hoque and R. W. Picard. Acted vs. natural frustration and delight: Many people smile in natural frustration. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011.

[31] L. A. Jeni, J. F. Cohn, and F. De La Torre. Facing imbalanced data–recommendations for the use of performance metrics. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013.

[32] D. W. Johnson et al. *Cooperative learning in the classroom*. ERIC, 1994.

[33] I. Kats-Gold, A. Besser, and B. Priel. The role of simple emotion recognition skills among school aged boys at risk of adhd. *Journal of Abnormal Child Psychology*, pages 363–378, 2007.

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

[35] K. Krippendorff. *Content analysis: An introduction to its methodology*. Sage, 2004.

[36] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 2010.

[37] A. Lim and H. G. Okuno. The mei robot: towards using motherese to develop multimodal emotional intelligence. *IEEE Transactions on Autonomous Mental Development*, 2014.

[38] G. C. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly. Automated measurement of children's facial expressions during problem solving tasks. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011.

[39] V. LoBue and C. Thrasher. The child affective facial expression (cafe) set: Validity and reliability from untrained adults. *Frontiers in psychology*, 2015.

[40] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010.

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

[42] S. Patel, K. R. Scherer, J. Sundberg, and E. Björkner. Acoustic markers of emotions based on voice physiology. In *Proceedings of the speech prosody*, 2010.

[43] J. Rehg, G. Abowd, A. Rozga, M. Romero, M. Clements, S. Sclaroff, I. Essa, O. Ousley, Y. Li, C. Kim, et al. Decoding children's social behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[44] J. Rong, G. Li, and Y.-P. P. Chen. Acoustic feature selection for automatic emotion recognition from speech. *Information processing & management*, 2009.

[45] B. E. Rossing and H. B. Long. Contributions of curiosity and relevance to adult learning motivation. *Adult Education Quarterly*, 1981.

[46] N. Sato and Y. Obuchi. Emotion recognition using mel-frequency cepstral coefficients. *Information and Media Technologies*, 2007.

[47] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[48] M. Suk and B. Prabhakaran. Real-time mobile facial expression recognition system–a case study. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2014.

[49] G. A. van Kleef. *The Interpersonal Dynamics of Emotion*. Cambridge University Press, 2016.

[50] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005.

[51] C.-H. Wu, Z.-J. Chuang, and Y.-C. Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 2006.

[52] X. Yu, F. Yang, J. Huang, and D. N. Metaxas. Explicit occlusion detection based deformable fitting for facial landmark localization. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013.

[53] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 2009.