A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity

De'Aira Bryant
dbryant@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Ayanna Howard
ayanna.howard@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

ABSTRACT

In recent news, organizations have been considering the use of facial and emotion recognition for applications involving youth such as tackling surveillance and security in schools. However, the majority of efforts on facial emotion recognition research have focused on adults. Children, particularly in their early years, have been shown to express emotions quite differently than adults. Thus, before such algorithms are deployed in environments that impact the wellbeing and circumstance of youth, a careful examination should be made on their accuracy with respect to appropriateness for this target demographic. In this work, we utilize several datasets that contain facial expressions of children linked to their emotional state to evaluate eight different commercial emotion classification systems. We compare the ground truth labels provided by the respective datasets to the labels given with the highest confidence by the classification systems and assess the results in terms of matching score (TPR), positive predictive value, and failure to compute rate. Overall results show that the emotion recognition systems displayed subpar performance on the datasets of children 's expressions compared to prior work with adult datasets and initial human ratings. We then identify limitations associated with automated recognition of emotions in children and provide suggestions on directions with enhancing recognition accuracy through data diversification, dataset accountability, and algorithmic regulation.

ACM Reference Format:
De'Aira Bryant and Ayanna Howard. 2019. A Comparative Analysis of Emotion-Detecting AI Systems with Respect to Algorithm Performance and Dataset Diversity. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3306618.3314284

1 INTRODUCTION

Understanding a child's emotional state is of great importance in numerous applications, from understanding levels of comfort when interacting with a therapy robot (Leo et al. 2015) to identifying degrees of engagement or feelings of frustration when interacting with virtual agents during a learning scenario (Littleworth et al. 2011). However, before intelligent systems can be deemed usable for these societal purposes, it is critical that we examine the validity of the systems used for emotion recognition and classification amongst children. In the emotion recognition domain, one of the requirements for validating the performance of any new classification algorithm is to evaluate it against established datasets. There has been valuable work on validating models for recognizing emotion constructed via machine learning in recent years; yet, this work has focused primarily on adults imaged in different lighting conditions, scales, and from various perspectives (Dupré et al. 2017, Stöckli et al. 20117, Bernin et al. 2017).

We have identified a gap in research with regard to validating models for emotion recognition in children. The first contribution of this paper is an in-depth comparison of publicly available datasets for research purposes that have conducted inter-rater reliability studies for validating the emotion labels associated with the facial expressions of children. Second, we have conducted an evaluation of eight commercially available emotion recognition systems against the five datasets of children expressions. To the best of our knowledge, this paper represents one of the few comparisons to be made on emotion recognition datasets and classification systems with a focus on children. We also highlight a rising concern with constructing classifiers for children while using validating datasets where children have poor representation. This challenge resonates with similar problems seen across the machine learning and artificial intelligence (AI) communities.

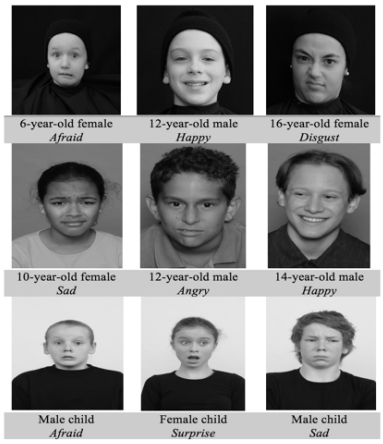

Figure 1: Example stimuli of children associated with the facial expression databases: Top: Dartmouth Database of Children 's Faces [8]; Middle: NIMH-ChEFS database; Bottom: Radboud Faces Database [11].

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314284

2 BACKGROUND & RELATED WORK

2.1 The Role of Emotions

The human face is an extremely complex source of insight into the inner-workings of the mind and body with the ability to express thousands of different facial configurations. Of these configurations, notable psychologist Paul Ekman found that there are six universal basic emotions: anger, disgust, fear, happiness, sadness, and surprise (Ekman 1992). These emotion classes, as interpreted from facial expressions, are key factors influencing social inter-human interaction. If AI agents are to be capable of navigating complex social scenarios with humans, it is critical that they are capable of perceiving these multiple emotion categories.

In addition to understanding the differences between the emotion categories and their implications, it is also necessary to consider the dynamic features that may affect certain subsets of the population. For example, as the bounds between emotion categories are traditionally socially constructed (Gordon, 1991), children often take several years to reach the levels of emotional intelligence that is often seen in adults (Durand et al. 2007, Mondloch et al. 2003). In turn, their expressions of specific emotions differ from adults in a variety of ways. For example, Saarni notes how children in their early years heavily associate emotions to facial expressions and therefore learn to express the concepts of happiness, sadness and anger earlier than the concepts of fear, surprise and disgust (Saarni 1999). As children have a limited amount of social emotional experiences, it can take many years for them to learn common social cues (Herba et al. 2006, Thomas et al. 2007).

2.2 Approaches to Emotion Recognition Systems

A majority of emotion recognition and classification systems utilize an approach based on the Emotional Facial Action Coding System (EmFACS) which encompasses mapping specific facial muscle configurations to the various emotional categories (Friesen and Ekman 2005). As described in (Dupré et al. 2017, Bernin et al. 2017), the general approach to classifying still images includes finding the face in the image, extracting the relevant features such as facial action units (AUs), and finally classifying the image using algorithms trained through various machine learning techniques. A non-exhaustive list of available emotion recognition systems, past and present, can be found in (Deshmukh and Jagtap 2017).

Although several efforts have relied on machines for recognizing emotions in children to enable their functionality, most have not done a systematic analysis of the performance of these emotion classification results in children. For example, in the realm of socially interactive robots, research robots use emotions to engage children in therapy or learning (Brown and Howard 2014, Metta et al. 2008, Simmons et al. 2003). However, their performance evaluation is based on measures of child engagement rather than on emotion recognition. In (Littleworth et al. 2011), accuracy measures were based on Action Units. Another research effort, (Khan, Meyer, and Bouakaz 2015), reported achieving a maximum overall recognition rate of 79% with the automated recognition of facial expressions for children when considering the full Dartmouth Database of Children Faces. The team later tested their classifier on the NIMH Child Emotional Faces Picture Set database and achieved a recognition rate of 68.4%. Although these efforts have begun to address some of the research gaps in validating models for emotion recognition in children, they have not evaluated these models against a variety of diverse datasets or considered performance metrics other than overall classification accuracy.

3 METHODOLOGY

Here, we introduce five image datasets comprised of the facial expressions of children. These datasets are available publicly for research purposes with labels and inter-rater reliability data provided. When assumptions had not been made in the past, we admitted into the study those images associated with an inter-rater reliability value of at least 75%. In the other cases, we applied the threshold values for inclusion used in the researchers'studies and published results. The five datasets compared were the NIMH Child Emotional Faces Picture Set (NIMH-ChEFS), the Dartmouth Database of Children's Faces, the Radboud Faces Database, the Child Emotions Picture Set (CEPS), and the Child Affective Facial Expressions Set (CAFE) (Figure 1). Next, we compare these systems through a diversity analysis and a comparison of human recognition rates on the images. We then introduce the eight selected emotion recognition systems and compare their various attributes.

	Fear	Anger	Disgust	Happiness	Sadness	Surprise
NIMH-ChEFS	96%	95%	-	99%	92%	-
Dartmouth	82%	87%	86%	96%	93%	88%
Radboud	86%	93%	89%	97%	90%	92%
CEPS	74%	88%	82%	96%	80%	81%
CAFE	70%	80%	75%	89%	79%	76%
Avg. Human Recognition	**82%**	**89%**	**83%**	**95%**	**87%**	**84%**

Table 1: Human ratings with good inter-rating reliability.

3.1 Children Facial Expression Datasets

The NIMH Child Emotional Faces Set (NIMH-ChEFS). This dataset contains images of the emotional faces of children ranging in age, ethnicity, and gender (Egger et al. 2011). The original picture set includes 534 pictures with 39 girls and 20 boys in the picture set (total N=59) covering 5 emotions (afraid, angry, happy, sad and neutral) and two gaze conditions (direct and averted). The child actors range in age from 10 to 17 years old with a mean age of 13.6 years old. Images are coded for emotion by a sample of 20 raters ranging in age from 22 to 70 (mean age 38.3). A cut-off point for inclusion was established of 15/20 (75%) of the raters correctly identifying the intended emotion, which excluded 52 pictures from the original set leaving a final set of 482 pictures.

The Dartmouth Database of Children's Faces. This dataset contains images of 40 male and 40 female Caucasian children ranging in age between 6 and 16 (Dalrymple, Gomez, and Duchaine 2013). The original picture set includes 1280 images covering 7 emotions (neutral, happy, sad, angry, afraid, surprise, and disgust).

The models photographed for the study ranged in age from 5 to 16 years old with a mean age of 9.72 years old. Images were coded for emotion by a random sample from 163 recruited adult raters. Each image was assessed by at least 20 raters for facial expression. For comparative analyses, we selected a cut-off point for inclusion of cases at 75% of the raters correctly identifying the intended emotion, which excluded 370 pictures from the original set leaving a final set of 910 pictures.

The Radboud Faces Database (RaFD). This dataset contains images of the emotional facial expressions of 4 male and 6 female Caucasian Dutch children (Langner et al. 2010). The original picture set includes 240 images covering 8 emotions (neutral, angry, sad, afraid, disgust, surprise, happy, and contempt). Images were coded for emotion by a random sample from 276 recruited raters with a mean age of 21.2; 238 were women. Each image was assessed by at least 20 raters for facial expression. We again established a cut-off point for inclusion at a threshold requiring at least 75% of the raters correctly identifying the intended emotion. This excluded 57 pictures from the original set leaving a final set of 183 pictures.

The Child Emotions Picture Set (CEPS). This dataset contains images of the emotional faces of Brazilian children ranging in age, ethnicity, and gender (Romani-Sponchiado et al. 2015). The picture set includes 273 pictures with 9 girls and 8 boys in the picture set (total N=17) covering 7 emotions (happy, sad, angry, disgust, afraid, surprise, and neutral) and 3 intensity levels. The children ranged in age from 6 to 11 years old with a mean age of 8.9 years old. Images were coded for emot ion by a sample of 30 psychologists as raters, with each image receiving at least 5 ratings. A cut-off point for inclusion was established by the researchers at 60% of the raters correctly identifying the intended emotion, which excluded 48 pictures from the original set leaving a final set of 225 pictures.

The Child Affective Facial Expressions Set (CAFE). This dataset contains images of the emotional faces of children ranging in age, ethnicity, and gender (LoBue and Thrasher 2015). The original picture set includes 1192 pictures with 90 girls and 64 boys in the picture set (total N=154) covering 7 emotions (happy, angry, sad, afraid, surprise, neutral, and disgust). Children range in age from 2 to 8 years old with a mean age of 5.3 years old. Images were coded for emotion by a sample of 100 raters. A cut-off point for inclusion was established by the researchers at 66% of the raters correctly identifying the intended emotion, which excluded 403 pictures from the original set leaving a final set of 789 pictures.

Table 1 summarizes the various emotional stimuli and the associated ratings that result when human raters are asked to label the basic emotions for each presented image.

3.2 Dataset Diversity

To break down the composition of the datasets, we introduce eight attributes of diversity that contribute to the makeup of image datasets used for emotion recognition. We use these metrics to derive a diversity rating for each dataset. This rating scale can be used to further illustrate the validity of new datasets and emotion recognition systems by assessing the diversity of the image data. The nine attributes contributing to the diversity rating include age, gender, ethnicity, gaze, geographic location of recruitment, clothing, pose, and number of emotion classes. We describe how we score values from 0 to 1 for each attribute in Table 2. We then show the scores for each of the five datasets in Table 3. A diversity rating of 1.0 is associated with a fully diverse dataset in terms of representation, setting and collection.

Diversity Metric	Rating Description
Age	Age diversity was ranked by considering how representative the dataset was of the desired population's age range. Given the range, a ratio was calculated for each age represented in the dataset compared to the ratio it would be if the dataset were equally distributed by age. If each ratio was above 90%, the dataset received a 1.
Gender	Gender diversity was ranked by considering the ratio of male to female children participants. The closer the ratio was to an equal distribution, the higher the rating.
Ethnicity	Ethnic diversity was ranked by considering whether the dataset consisted of children from a single ethnic background or multiple. If only children from a single ethnic background existed in the dataset, the dataset received a 0. Otherwise, the dataset received a 1.
Gaze Direction	Gaze direction was ranked by considering whether the images were taken with multiple gaze directions. NIM-ChEFS was the only dataset to have a diverse selection of images with different gaze directions.
Geographic Region	Geographic region was ranked by considering the dataset's collection process. Each of the datasets recruited children from a single geographic region and were therefore granted a score of 0.
Clothing	Clothing was ranked by considering the diversity of clothing shown in the images. Datasets where children all wore identical outfits received a score of 0. Otherwise, datasets received a 1.
Pose	Pose was ranked by considering whether the images consisted of entirely staged images or not. Ceps was the only dataset which included spontaneous emotional expressions.
Num. Classes	The classes metric was ranked by considering the number of emotion classes that existed in the dataset. If a dataset included at least the six basic emotions, it was ranked with a 1. If not, the ratio of classes to the 6 basic emotions was used as the score.

Table 2: Metrics used for scoring and assessing the diverse makeup of an image dataset used for emotion recognition.

	Age	*Gen.*	*Ethn.*	*Gaze*	*Geo.*	*Clo.*	*Pose*	*Classes*	SUM	Diversity Rating
NIM_ ChEFS	1	0.51	0.33	1	0	1	0	0.67	4.51	**0.56**
Dartmouth	1	1	0.17	0	0	0	0	1	3.17	**0.40**
Radboud	1	0.67	0.17	0	0	0	0	1	2.84	**0.36**
CEPS	1	0.88	0.5	0	0	1	1	1	5.38	**0.67**
CAFE	1	0.71	0.67	0	0	0	0	1	3.38	**0.42**

Table 3: Dataset diversity rating breakdown for the 5 datasets of children emotion expression. Gender is abbreviated by "gen", Ethnicity: "ethn", Clothing: "clo", and Geography: "geo".

Matching Score (TPR)							
	Happy	Sad	Fear	Disgust	Anger	Surprise	AVG. MS
Google	99.47%	52.39%	-	-	26.32%	89.20%	66.84%
Sighthound	91.39%	50.80%	52.81%	39.37%	60.37%	77.70%	62.07%
Face++	91.56%	59.84%	19.14%	55.46%	48.97%	91.99%	61.16%
Amazon	98.42%	27.66%	-	10.06%	27.52%	89.90%	50.71%
Microsoft	99.30%	66.76%	16.50%	36.31%	48.74%	86.41%	59.00%
Skybiometry	76.94%	28.19%	49.50%	74.64%	31.33%	84.90%	57.58%
Affectiva	94.52%	23.17%	8.88%	64.75%	11.14%	90.91%	48.90%
Kairos	51.68%	18.55%	15.09%	30.12%	80.44%	65.06%	43.49%

Figure 2: Matching Scores (TPRs) for each emotion recognition system categorized by each emotion category. Fear and disgust images were not considered for the Google Vision API and fear images were not considered for Amazon Rekognition as these two systems do not provide confidence values for those emotions. Systems are listed in order of highest to lowest average matching scores.

3.3 Emotion Recognition Systems

To evaluate the performance and limitations of AI-based emotions recognition systems, we selected emotion recognition systems that had either an API or SDK which allowed the emotion recognition capabilities to be embedded into other applications. After a systematic review of the field, we included eight systems in our analysis: Affectiva, Google Vision API, Microsoft Emotion API, Amazon Rekognition, Face++, Kairos, Sighthound, and Skybiometry.

Commercially, these systems are being used in a variety of applications ranging from academic research, to advertising, to hospitality, to retail, to education, etc. Some have already been embedded into a variety of everyday technology. As such, there are potential impacts on many diverse groups in the world, including children. This work seeks to analyze the efficacy of these emotion recognition systems by assessing their performance on a variety of children emotion datasets, allowing us to visualize their usage potential in real-world scenarios involving youth.

4 PROCEDURE

We utilize a similar approach described in (Bernin et al. 2017) where we conduct a black box test for each emotion recognition system (Patton 2006). We first store the ground truth labels of the images. Next, we process each of the images from each of the datasets through each of the emotion recognition systems. We then normalize the results for an equal comparison. A maximization function is then used to determine the emotion label with the highest confidence value. Finally, we compare the system-produced predicted label to the ground truth label and store the results [1].

Google Vision API did not offer confidence values for the emotions of fear and disgust. Amazon Rekognition did not offer confidence values for the emotion fear. To assess these two algorithms fairly, we did not include the ratings for images labeled as emotions they do not provide confidence intervals for in the results section below.

[1]These black box tests occurred progressively between May 2018 and July 2018. As these systems have regular updates and changes to algorithmic functionality, it is possible that these results could differ if obtained at a later time.

5 RESULTS

We analyze the results of the emotion recognition systems by assessing the matching scores (TPR), positive predictive values, and failure to compute (FTC) rates of the data.

5.1 Matching Score (True Positive Rate)

Matching score, also known as accuracy, sensitivity or true positive rate, gives insight into how much of a particular class an emotion recognition system can accurately classify. Matching score is defined as the ratio between the number of true positives to the total number of total actual positives. True Positives represents the number of images where the predicted emotion label matches the ground truth label and total Actual Positives represents the total number of images with the ground truth emotion label. The matching scores for each emotion and each system can be seen in Figure 2.

5.2 Positive Predictive Value

Positive predictive value (PPV) gives insight into how much trust can be placed in a recognition system to assess a particular label. It is a measure of how often the predicted class is actually the ground truth. The formula for PPV can be seen in (1) and PPV scores for each emotion and each system can be seen in Figure 3.

$$PPV = \frac{(MS * prevalence)}{(MS * prevalence) + (1 - specificity) * (1 - prevalence)} \quad (1)$$

where PPV is positive predictive value and MS is matching score. Prevalence is defined as the ratio of the total Actual Positives to the total number of images classified. Specificity is defined as the ratio of the True Negatives to the total Actual Negatives.

5.3 Failure to Compute (FTC) Rate

A prerequisite to facial emotion classification is facial detection. There were some instances where the systems could not identify a face in an image and therefore would not provide emotional data.

Positive Predictive Value							
	Happy	Sad	Fear	Disgust	Anger	Surprise	Avg. PPV
Microsoft	67.83%	84.80%	100.00%	94.74%	76.34%	56.49%	**80.03%**
Google	57.95%	68.17%	-	-	97.46%	89.51%	**78.27%**
Sighthound	83.07%	87.61%	64.52%	74.46%	62.98%	66.37%	**73.17%**
Face++	78.35%	71.43%	61.05%	67.25%	72.79%	54.32%	**67.53%**
Skybiometry	93.38%	86.18%	65.07%	46.33%	52.85%	45.81%	**64.94%**
Kairos	94.25%	72.73%	37.07%	59.51%	24.81%	68.63%	**59.50%**
Amazon	51.86%	54.45%	-	62.50%	54.79%	52.65%	**55.25%**
Affectiva	74.92%	54.68%	44.23%	35.50%	54.17%	34.33%	**49.64%**

Figure 3: Positive Predictive Value rates for each emotion recognition system categorized by each emotion category. Fear and disgust images were not considered for the Google Vision API and fear images were not considered for Amazon Rekognition as these two systems do not provide confidence values for those emotions. Systems are listed in order of highest to lowest average PPV.

We use FTC rates to illustrate how often this scenario occurred. Face++, Google, Microsft, Amazon and Sighthound all had FTC rates less than 1%. Skybiometry, Kairos, and Affectiva each had FTC rates of 2.39%, 9.15%, and 15.34% respectively.

6 DISCUSSION

Our results indicate that the emotions of happiness and surprise were most easily identified and classified correctly by each of the emotion recognition systems, except for Kairos. Fear and sadness were amongst the hardest to identify and classify. Google's Vision API had the highest average matching scores for the images it processed, which excluded images labeled as fear and disgust. Sighthound had the next highest overall matching scores. Face++, Microsoft Emotion API and Skybiometry ranked very closely to Sighthound in terms of matching scores.

In terms of PPV, Microsoft's Emotion API produced the best overall results with 100%, 95%, and 85% PPV rates for fear, disgust, and sadness respectively. Google's Vision API came in a close second with the highest PPVs for anger and surprise. An interesting observation can be observed when comparing the PPV rates to the matching scores. For example, Microsoft's Emotion API has a 100% PPV rate and a 16.5% matching score for fear. This shows that though the recognition system only produced the fear label for images with a ground truth label of fear, the system only picked up a small fraction of the images with that label. This trend was observed across multiple systems in our analysis. This illustrates the importance of the threshold values potentially used within each of the recognition systems. Our hypothesis is that systems with tighter thresholds for classification tend to have higher PPV rates whereas systems with looser thresholds tend to have higher TPR rates.

An interesting question arises when considering which metric should be held of highest importance. Should an emotion recognition system aim to classify the most instances of a particular category? Or, should a system aim to maximize the confidence in its predictive value? Should users of the technology be able to have a say based on their intended application? Is there a way to best maximize the two using additional input parameters? These are questions that the creators of such technology must consider in future iterations of their software.

Additionally, a similar comparative analysis using adult emotion image datasets found that Sighthound and Microsoft's Emotion API had an average 76.1% and 61.3% matching score respectively (Dehghan et al. 2017). For Microsoft, this is comparable to the 59% average score for the children emotion datasets. However, Sighthound performed worse on children's faces than adult faces, with only 62.07%. Affectiva reports that their system achieves accuracy in the high 90th percentile for key emotions, yet their average matching score for the children datasets was 48.9%, among the worst of the analyzed systems. Affectiva also had the lowest average PPV and a failure to compute rate of about 15%.

Human accuracy of the selected images for analysis had accuracy rates for each basic emotion above 80%. No system had this level of performance on more than two of the six emotions. These results provide further evidence that popular emotion recognition systems have not thoroughly considered children as a part of their target population. Yet, there is little to no regulation on what categories of people this software can or cannot be used for. With the psychological differences in the expression of emotions found in children, it is critical to develop either improved, standalone or adaptive emotion recognition software to adequately service this youthful audience.

7 CONCLUSION

In this work, we assess and evaluate five datasets of children emotional expression and eight emotion recognition systems. We first evaluate the composition of the different datasets through a diversity rating which considers the attributes of age, gender, ethnicity, gaze, geographic region, clothing, pose, and number of classes. Next, we evaluate the human performance of recognition between the various datasets. Finally, we conduct a comparative analysis between the eight emotion recognition systems using the five datasets. From this analysis, we conclude that most systems performed worse when compared to human raters and similar studies conducted using adult emotional data.

As we have seen, the least recognized emotion among the human raters was fear. This poorer recognition rate is also reflected by

the various emotion recognition systems. Given that the biases in recognition rate for human raters seems to also be reflected in the various systems, there is a concern that these algorithms are reflecting some degree of human biases. Recently, there has been an upsurge of attention given to machine learning algorithms and the practices of inequality and discrimination that are potentially being built into them (Buolamwini and Gebru 2018, Crawford 2016). We know that imbalances exist in training sets. There is a danger that specific imbalances in the training data will result in biases that may be implicit and unrecognized. Additional work is needed to address these issues in algorithmic learning and classification.

Additionally, recent articles, (Lapowsky 2018, Vanderklippe 2018), detail how facial and emotion recognition technology is being considered for educational environments in an attempt to target societal issues of student surveillance and security. The results of this work demonstrate that these potential applications are undeniably premature. This is an immediate and pressing problem. If these systems are not holistically designed for the audiences in which they are inevitably impacting, we will continue to see the perpetuation of implicit bias and unfairness in these systems with potentially devastating impacts.

We see the development of best practices for directly targeting these issues surrounding bias and inclusion with establishing representativeness in training sets, evaluating the validity of datasets for testing procedures, and calling for some third-party oversight for the inclusion of recognition, classification, and recommender systems that are to be used in societal applications. With these recommendations, technology can continue to mature progressively and without the taint of inherent human biases. These precautionary enhancements will pave the way for future affective technology and allow for a variety of useful applications in their due time.

ACKNOWLEDGMENTS

This research is based upon work partially supported by funding from the Linda J. and Mark C. Smith Endowed Chair, the NSF-GRFP under Grant No. DGE-1650044 and NSF Award No. 1849101.

REFERENCES

[1] Affectiva, Affectiva, Inc. https://www.affectiva.com/.
[2] Amazon Rekognition, Amazon, https://aws.amazon.com/rekognition/.
[3] Bernin, A., Müller, L., Ghose, S., von Luck, K., Grecos, C., Wang, Q., & Vogt, F. 2017. Towards more robust automatic facial expression recognition in smart environments. In Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments. (pp. 37-44). ACM.
[4] Brown, L. and Howard A. 2014. Gestural Behavioral Implementation on a Humanoid Robotic Platform for Effective Social Interaction. IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN), pp. 471 – 476.
[5] Buolamwini, J. and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency, pp. 77-91.
[6] Crawford, K. 2016. Artificial Intelligence's White Guy Problem, New York Times - Opinion, http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html.
[7] Dalrymple KA, Gomez J and Duchaine B. 2013. The Dartmouth Database of Children's Faces: Acquisition and Validation of a New Face Stimulus Set. Urgesi C, ed.PLoS ONE. 2013;8(11).
[8] Dehghan, A., Ortiz, E. G., Shu, G., & Masood, S. Z. 2017. Dager: Deep age, gender and emotion recognition using convolutional neural network. arXiv preprint arXiv:1702.04280.
[9] Deshmukh, R. S., & Jagtap, V. 2017. A survey: Software API and database for emotion recognition. In Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on (pp. 284-289). IEEE.
[10] Dupré, D., Andelic, N., Morrison, G., & McKeown, G. 2017. Accuracy of three commercial automatic emotion recognition systems across different individuals and their facial expressions. In 2018 IEEE International Conference on Pervasive Computing and Communications: Proceedings IEEE.
[11] Durand, K., Gallay, M., Seigneuric, A., Robichon, F., & Baudouin, J. Y. 2007. The development of facial emotion recognition: The role of configural information. Journal of experimental child psychology, 97(1), 14-27.
[12] Egger HL, Pine DS, Nelson E, et al. 2011. 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.20(3):145-156.
[13] Ekman, P. 1992. An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.
[14] Face++, Megvii Technology, https://www.faceplusplus.com/.
[15] Friesen, W.V, Ekman, P, EMFACS-7: Emotional Facial Action Coding System, Unpublished manuscript, University of California at San Francisco, 1983.
[16] Google Vision API, Google Cloud Platform, https://cloud.google.com/vision/.
[17] Gordon, S. L. 1991. The socialization of children's emotions: emotional culture, competence, and exposure. Children's understanding of emotion, 319.
[18] Herba, C. M., Landau, S., Russell, T., Ecker, C., & Phillips, M. L. 2006. The development of emotion–processing in children: Effects of age, emotion, and intensity. Journal of Child Psychology and Psychiatry, 47(11), 1098-1106.
[19] Howard, A, Zhang, C., and Horvitz, E. 2017. Addressing Bias in Machine Learning Algorithms: A Pilot Study on Emotion Recognition for Intelligent Systems. IEEE International Workshop on Advanced Robotics and its Social Impacts, Austin, TX, March.
[20] I. Lapowsky. 2018. Schools Can Now Get Facial Recognition Tech For Free. Should They? https://www.wired.com/story/realnetworks-facial-recognition-technology-schools.
[21] Kairos, Kairos AR, https://www.kairos.com/.
[22] Khan, R.A., Meyer, A., Bouakaz, S. 2015. Automatic Affect Analysis: From Children to Adults, International Symposium on Visual Computing, ISVC 2015, 304-313.
[23] Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Hawk, S.T., & van Knippenberg, A. 2010. Presentation and validation of the Radboud Faces Database. Cognition & Emotion, 24(8).
[24] Leo, M., Coco, M. D., Carcagnì, P., Distante, et al. 2015. Automatic Emotion Recognition in Robot-Children Interaction for ASD Treatment., in 'ICCV Workshops', IEEE, pp. 537-545 .
[25] Littleworth, G., Bartlett, M.S., Salamanca, L.P. and Reilly, J. 2011. Automated Measurement of Children's Facial Expressions during Problem Solving Tasks. IEEE Int. Conference on Automatic Face and Gesture Recognition, pp. 30–35.
[26] LoBue, V., & Thrasher, C. 2015. The Child Affective Facial Expression (CAFE) set: validity and reliability from untrained adults. Frontiers in psychology, 5, 1532.
[27] Metta, G., Sandini, G. Vernon, D. Natale, L. Nori, F. 2008. The iCub humanoid robot: an open platform for research in embodied cognition, 8th workshop on performance metrics for intelligent systems, pp. 50-56.
[28] Microsoft Emotion API, Microsoft Azure, https://azure.microsoft.com/en-us/services/cognitive-services/emotion/.
[29] Mondloch, C. J., Geldart, S., Maurer, D., & Le Grand, R. 2003. Developmental changes in face processing skills. Journal of experimental child psychology, 86(1), 67-84.
[30] Patton, R. 2006. Software testing. Pearson Education India. pp.55.
[31] Romani-Sponchiado, A., Sanvicente-Vieira, B., Mottin, C., Hertzog-Fonini, D.,& Arteche, A. 2015. Child Emotions Picture Set (CEPS): Development of a database of children's emotional expressions. Psychology & Neuroscience, 8(4), 467.
[32] Saarni, C. 1999. The development of emotional competence. Guilford Press.
[33] Sighthound, Sighthound, Inc., https://www.sighthound.com/.
[34] Simmons R., et al. 2003. GRACE: An Autonomous Robot for the AAAI Robot Challenge, AI Magazine, Vol. 24(2), pp. 51-72.
[35] Skybiometry, Skybiometry UAB, https://skybiometry.com/.
[36] Stöckli, S., Schulte-Mecklenbeck, M., Borer, S., & Samson, A. C. 2017. Facial expression analysis with AFFDEX and FACET: A validation study. Behavior research methods, 1-15.
[37] Thomas, L. A., De Bellis, M. D., Graham, R., & LaBar, K. S. 2007. Development of emotional facial recognition in late childhood and adolescence. Developmental science, 10(5), 547-558.
[38] Vanderklippe, N. 2018. In China, Classroom Cameras Scan Students Faces for Emotion, Stroking Fears of New Form of State Monitoring. https://www.theglobeandmail.com/world/article-in-china-classroom-cameras-scan-student-faces-for-emotion-stoking/.