Training Affective Computer Vision Models by Crowdsourcing Soft-Target Labels

Peter Washington1 · Haik Kalantarian2 · Jack Kent2 · Arman Husic2 · Aaron Kline2 · Emilie Leblanc2 · Cathy Hou3 · Cezmi Mutlu4 · Kaitlyn Dunlap2 · Yordan Penev2 · Nate Stockham5 · Brianna Chrisman1 · Kelley Paskov6 · Jae-Yoon Jung2 · Catalin Voss3 · Nick Haber7 · Dennis P. Wall8

Received: 26 March 2021 / Accepted: 12 September 2021 / Published online: 27 September 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
Emotion detection classifiers traditionally predict discrete emotions. However, emotion expressions are often subjective, thus requiring a method to handle compound and ambiguous labels. We explore the feasibility of using crowdsourcing to acquire reliable soft-target labels and evaluate an emotion detection classifier trained with these labels. We hypothesize that training with labels that are representative of the diversity of human interpretation of an image will result in predictions that are similarly representative on a disjoint test set. We also hypothesize that crowdsourcing can generate distributions which mirror those generated in a lab setting. We center our study on the Child Affective Facial Expression (CAFE) dataset, a gold standard collection of images depicting pediatric facial expressions along with 100 human labels per image. To test the feasibility of crowdsourcing to generate these labels, we used Microworkers to acquire labels for 207 CAFE images. We evaluate both unfiltered workers and workers selected through a short crowd filtration process. We then train two versions of a ResNet-152 neural network on soft-target CAFE labels using the original 100 annotations provided with the dataset: (1) a classifier trained with traditional one-hot encoded labels and (2) a classifier trained with vector labels representing the distribution of CAFE annotator responses. We compare the resulting softmax output distributions of the two classifiers with a 2-sample independent \( t \)-test of L1 distances between the classifier’s output probability distribution and the distribution of human labels. While agreement with CAFE is weak for unfiltered crowd workers, the filtered crowd agree with the CAFE labels 100% of the time for happy, neutral, sad, and “fear + surprise” and 88.8% for “anger + disgust.” While the F1-score for a one-hot encoded classifier is much higher (94.33% vs. 78.68%) with respect to the ground truth CAFE labels, the output probability vector of the crowd-trained classifier more closely resembles the distribution of human labels \( (t = 3.2827, p = 0.0014) \). For many applications of affective computing, reporting an emotion probability distribution that accounts for the subjectivity of human interpretation can be more useful than an absolute label. Crowdsourcing, including a sufficient filtering mechanism for selecting reliable crowd workers, is a feasible solution for acquiring soft-target labels.

Introduction

Machine learning models which predict human emotion from images of facial expressions are increasingly used in interactive systems [1–5] and applications such as multimodal sentiment analysis [6–8], healthcare [9, 10], and autonomous vehicles [11]. Emotion recognition is traditionally modeled as a classification problem, where the model predicts a discrete emotion category. However, facial expressions are often ambiguous [12–14], and it is often not ideal for a machine learning model to output a single class for a subjective label. Fortunately, most supervised learning methods output a probability distribution over all possible classes. Sometimes, the affective computing system will visualize this distribution to the user [2]. Examples include commercial emotion detection services like Affectiva [15, 16] and autonomous vehicle displays [5]. In a large number of use cases, however, only the class with the highest probability is visualized [3, 4].

While the paradigm of training a model with a discrete one-hot encoded label and predicting a probability distribution is reasonable when the training data have indisputable
labels, images of facial expressions can have ambiguous labels or even multiple correct labels simultaneously, and the label should ideally represent this inherent uncertainty. Soft-target labeling, where the training labels represent a probabilistic distribution rather than a one-hot encoded label, is an established solution to this issue. Training with soft-target labels results in classifiers which predict probability distributions representative of the soft-target labels [17–19]. We hypothesize that crowdsourcing can generate distributions which mirror those generated in a lab setting.

Here, we explore the use of crowdsourcing to acquire a distribution of labels for images with ambiguous or multiple classes (we call these “subjective labels”). We first describe the acquisition of crowdsourced labels for four representative images which we display to the reader along with the distribution of crowd responses to demonstrate the phenomena of subjective labels in affective computing. We then crowdsource the labeling of a subset of the Child Affective Facial Expression (CAFE) dataset, a collection of emotive images of children which conveniently comes with 100 independent human annotations per image. We next show that the crowdsourced distribution mirrors the original CAFE distribution, validating the feasibility of crowdsourcing for generating a reliable and representative distribution of human labels for an image. Finally, we compare the performance of two versions of a convolutional neural network (CNN) trained on CAFE: one with traditional one-hot encoded vectors and the other with soft-target labels based on CAFE annotator responses. We find that the classifier trained with soft targets results in classifier predictions that much more closely mirror the true human distribution on independent subjects not included in the training set. We hope that this work will be of use to designers and developers of machine learning models for affective computing systems who wish to provide probabilistic outputs to the end user.

**Related Work**

While crowdsourcing and soft-target labels have been studied in affective computing, we are the first to explore the feasibility of using crowdsourcing to acquire reliable soft-target labels for computer vision emotion detection. We describe related work below.

**Facial Emotion Detection**

Facial emotion detection is a key challenge for machine learning. For intelligent machines to convincingly pass the Turing test [20], an understanding of human emotion is crucial. There has been a strong body of machine learning literature for detecting human affect from a variety of data streams, including audio [7, 8, 21], text [1, 22], images [23], and video [24–26]. Here, we focus on image-based emotion detection from facial expressions.

Fundamental to a successful computer vision approach for affective computing is the feature representation of the image, and there are several approaches to engineering such features. A common approach is to extract facial keypoints and use a feature representation consisting of the coordinates of the keypoints [24, 27–29]. This approach works well when the dataset is small, as the representation itself is compact and therefore amenable to lightweight learning approaches such as logistic regression, support vector machines, and decision trees. Another feature extraction approach, CNNs, can automatically learn relevant nonlinear feature maps. CNNs oftentimes result in superior performance to other methods when the dataset is sufficiently large [30, 31].

**Emotion Detection with Subjective Labels**

Paul Ekman posited that there are seven fundamental human emotions which are universal across cultures and geographic boundaries: happy, sad, surprise, anger, fear, disgust, and contempt [32, 33]. However, these expressions are not mutually exclusive. Du et al. discussed the existence of compound emotions, or combinations of existing emotions to form new ones [34]. Examples of compound emotions include “happily surprised”, “fearfully surprised,” and “fearfully disgusted.” Through smartphone sensing, Zhang et al. found that pairs of emotions which are often presented simultaneously include (happy, surprised), (sad, disgust), and (sad, fear) [35]. This issue has been explored for emotional speech [12, 14]. While some emotions may be jointly expressed, others may be singular yet ambiguous. The issue of subjectivity in training labels, whether due to ambiguous labels or multiple correct labels, has been documented in the fields of digital health and affective computing in particular [13, 36–39].

The topic of subjective labels in affective computing datasets containing speech and audio data has been explored in prior work. Mower et al. represent emotion labels at the granularity of utterances, thereby representing a time profile of how the dominant emotion in speech changes quickly over time [40, 41]. Fujioka alternates between updating neural network parameters and updating sample importance parameters at each training iteration [42]. Ando et al. utilize soft-target training, where the emotion labels are based on the proportion of human annotations instead of the traditional one-hot encoding [17].

Soft-target training is a general machine learning method for handling subjective training labels. This approach is particularly desirable when multiple labels are acquired per image. Classification of soft labels can be beneficial because they can account for inherent subjectivity in labels.
and are robust against random noise [43]. A variation of this method is a soft loss function, which consists of subtracting the minimum between-class distance from the maximum within-class distance [44]. Soft-target and loss training have been shown to outperform hard target training (one-hot encoding) when the training goal is to produce an output distribution like the distribution of annotator labels [45], and this phenomenon has been observed across several datasets and tasks [46].

The issue of subjective labels has also been explored in multimodal sentiment analysis, where the goal is to predict sentiment from multiple data streams [6–8], including affect-enriched videos. Chaturvedi et al. created a fuzzy classifier for predicting the degree to which several emotions are expressed in a particular image [47]. Another approach is to predict the amount of valence and arousal displayed on continuous axes (regression) rather than predicting categories (classification) [48–52].

Crowdsourcing with Subjective Labels

There are several bodies of work which describe approaches to handling crowdsourced labels. Kairam and Heer hypothesize that there are intrinsic but valid differences between crowd workers when labeling data points and therefore categorize workers by their labeling patterns [53]. Other examples of categorizing workers include measuring sample informativeness, active and cooperative learning strategies, and controlling for labeler trustworthiness metrics [38].

There have been other statistical learning techniques beyond the soft-target labeling discussed above which have been successful with crowdsourced labels. Rodrigues and Pereira add an extra “crowd layer” at the end of a traditional CNN architecture trained to predict the outputs of each labeler individually and therefore the biases of crowd workers [54].

Crowdsourcing has been used to acquire emotion labels of images. Korovina et al. found that crowd workers labeling discrete emotion categories on a color wheel had low agreement scores (Kappa value less than 0.15) [55] while consistency between workers when labeling valence and arousal was much stronger [56].

Methods

Acquiring Crowd Labels for CAFE Images

We use the CAFE dataset [57, 58], which is the largest public dataset of front-facing images depicting children emoting. CAFE is used as a benchmark in several affective computing publications [57, 58] and is the standard evaluation dataset for pediatric affective computing. CAFE was originally labeled by 100 untrained human raters, and the raw distribution of 100 human labels per image is provided along with the ground truth labels. For example, the first image in the dataset was labeled as “angry” by 62% of raters and as “disgusted” by 25% of raters. All other emotions received 5% or fewer labels. One could hypothesize from these numbers that the image looks “mostly angry” with “some disgust.” Manually inspecting the image reveals a facial expression which could reasonably be categorized as either “angry” or “disgusted” depending on the context. (CAFE images are protected by copyright and cannot be republished, so we refer to this image by its filename in the publicly available dataset: F-AA-01_052-Angry.jpg).

To validate the capability of crowdsourcing to produce a reliable ground truth label distribution, we crowdsourced the task of labeling CAFE images and compared the resulting crowd-generated distribution to the distribution reported in the original CAFE dataset. All crowdsourcing was conducted on Microworkers.com, a crowdsourcing platform similar to Amazon Mechanical Turk [59] but with a more globally representative pool of workers [60]. Each task consisted of labeling one of seven emotion categories (happy, sad, surprised, angry, fearful, disgusted, and neutral) for a subset of images in CAFE. We chose to limit rater labels to absolute ratings (one-hot representations) because we wanted to capture the relative weighting of each emotion within an image. Because humans are notoriously poor at precisely quantifying relative contributions of individual components in mixed representations, especially in the case of human emotion recognition [61–63], we asked each rater to only provide the most salient emotion according to their interpretation. By acquiring labels from 100 independent crowd workers per image, each providing their vote for the most prominent emotion, we created a representation describing the between-subject subjectivity of the emotion expressed in the image.

We acquired labels for 131 randomly selected images from café, and we solicited 100 crowd labels per image. We manually checked each label for correctness, and workers with consistently high-quality labels were recruited for additional labeling tasks for 76 separate images. Here, “high quality” means that the authors could potentially agree with the label (e.g., a “happy” label for a clearly “sad” image would not be accepted, but a “fearful” label for a “fearfully surprised” expression would be accepted). Our goal when excluding workers without consistently high-quality labels was to filter out crowd workers who were answering randomly to receive payment, as this is a common issue in crowdsourcing [64–67]. We analyzed both the filtered and unfiltered worker labels on different sets of CAFE images to measure the possibility that filtering workers could mask ambiguity of the labels.
All crowdsourcing tasks were approved by the Institutional Review Board (IRB) of Stanford University. All workers were required to sign an electronic consent form approved by the IRB before participating in the task.

**Training and Testing with Crowd Probability Distributions**

Traditionally, multi-class models are trained with categorical cross-entropy loss, where \(\Sigma\) is the summation operator, \(C\) is the number of classes, \(p_i\) is the ground truth probability of class \(i\), and \(q_i\) is the classifier prediction for class \(i\):

\[
-\sum_{i=1}^{C} p_i \log(q_i)
\]

When the true classes are indisputable, which is the usual assumption for classification, then the ground truth probability distribution \(p_i\) is a one-hot encoding (i.e., a probability of 1 for the “true” class and a probability of 0 for all other classes). In the case of subjective classes where the true label may consist of a weighted combination of multiple classes, like in emotion datasets where complex emotions are present, we hypothesize that providing soft-target labels instead of one-hot encodings will result in classifier predictions for separate human subjects which resemble the human annotator response distribution.

We trained a machine learning model using two sets of image labels: (1) the original CAFE labels as one-hot encoded vectors and (2) soft-target vectors representing the distribution of 100 human responses from the original CAFE dataset. We held out all images from 5 randomly selected child subjects from CAFE (F-AA-01, F-EA-39, M-LA-08, M-AA-11, and F-LA-13, corresponding to one female African American, one female European American, one male Latin American, one male African American, and one female Latin American) and used these as test set images. The rest of the images were used to train the classifier. A total of 1141 images (196 angry, 180 disgusted, 135 fearful, 206 happy, 222 neutral, 103 sad, and 99 surprised) were used in the train set, and 51 images (9 angry, 11 disgusted, 5 fearful, 9 happy, 8 neutral, 5 sad, and 4 surprised) were used in the test set.

We transfer learned on a ResNet-152 [68] CNN pretrained on ImageNet [69]. We trained each neural network using the Keras framework [70] with a TensorFlow [71] backend for 100 epochs with a batch size of 16 and a learning rate of 0.0003 using Adam optimization [72]. To increase generalization of the training process and reduce overfitting, we applied the following data augmentation strategies: a rotation range of 7 degrees, a zoom range of 15%, a shear range of 5%, a brightness range of 70 to 130%, and horizontal flipping.

**Results**

**Demonstration of Subjective Emotions**

The methods described here are not specific to CAFE. We focus on CAFE in this paper as a case study of a popular affective computing dataset and as a dataset which provides ground truth labels for many human annotators (100) per image. However, CAFE images are subject to copyright and cannot be republished. To provide the reader with visual examples of facial expressions with large numbers of crowd annotations per image, we display free-to-republish images in Fig. 1. For each image, we acquired 200 crowdsourced labels from Microworkers.com, as described above.

Figures 1A shows an image that could be labeled as either angry or disgusted, and Fig. 1C shows an image that is possibly angry, fearful, surprised, or some combination of the 3. Further context is required to reach full confidence about the true classes. Figure 1B displays a compound emotion, where the individual appears to be “fearfully surprised.” Assigning only a single category to the image would be misleading. Figure 1D depicts a situation where it is unclear whether the individual’s neutral face looks sad or if that individual is making a sad face. In cases like this, a personalized emotion recognition model would likely be required.

We also quantified the subjectivity of images in CAFE. We measured the number of images with 80% of annotations represented with the top-N most frequent labels for N ranging from 1 to 5 inclusive (histogram in Fig. 2). We see that while many emotions do not contain much subjectivity (\(N = 1\)), most images are either ambiguous between or compound with 2 or more emotions. When the cutoff is increased to 90% (Fig. 3), the number of subjective labels increases further.

**Comparison of CAFE Labels and Crowd Performance**

When looking at the majority consensus label, the filtered crowd agreed with the CAFE labels 100% of the time for happy, neutral, sad, and surprise. There was 90% agreement for disgust, 75% agreement for anger, and 50% agreement for fear. When combining commonly confused labels into one class (anger + disgust and fear + surprise), the filtered crowd agreed with the CAFE labels 100% of the time for happy, neutral, sad, and “fear + surprise and 88.8% for anger + disgust.

By contrast, the unfiltered crowd workers did not agree as strongly with the CAFE labels when looking at the majority consensus, highlighting the need for quality
Tables 1 and 2 compare the distribution of labels from the original CAFE labelers as well as the filtered and unfiltered crowd workers (respectively) for a single image. In both the filtered and unfiltered cases, the distributions qualitatively mirror each other in terms of their peaks. In all cases, peaks which appear in the CAFE annotator distribution also appear...
in the crowd distribution, and vice versa. However, these distributions are noisy, and the relationship between the peaks cannot be guaranteed (e.g., if anger has more labels than disgust for CAFE annotators, disgust may have more ratings for crowd annotators). The generated crowd distributions must therefore be regarded as a noisy approximation to the true probability distribution, and further work should account for this noise in the label representation.

**Training and Testing with Crowd Probability Distributions**

We evaluate the models with F1-score rather than accuracy because CAFE is not a balanced dataset. When training with the one-hot encoded labels, the F1-score on the held-out test set is 94.33%. We emphasize that this high performance is misleading due to the ambiguity of the ground truth labels. When training with vectors representing the distribution of human labels, the F1-score on the held-out test-set is 78.68%. While the F1 score is lower when training with human distribution labels, the distribution of emotion predictions much more closely resembles the distribution of human labels for the distribution-trained classifier. For many applications of affective computing, having a representative label distribution is more important than absolute accuracy. The mean L1 distance between the human label distribution for the test set and distribution-trained classifier is 0.3727 (SD = 0.3000); the mean L1 distance between the human label distribution and one-hot encoding-trained classifier is 0.6078 (SD = 0.4143). The difference in L1

| Image | CAFE labeler distribution (count) | Crowdsourced labeler distribution (count) |
|-------|----------------------------------|------------------------------------------|
| 9990-angry_F-AA-15.jpg | 30, 37, 15, 8, 0, 8, 2 | 7, 3, 0, 4, 0, 0, 0 |
| 10108-angryopen_F-AA-15.jpg | 29, 6, 35, 1, 1, 23, 5 | 2, 2, 4, 0, 0, 6, 0 |
| 10194-disgust_F-AA-15.jpg | 3, 86, 3, 2, 1, 5, 0 | 2, 10, 0, 0, 0, 2, 0 |
| 10288-disgustwithtongue_F-AA-15.jpg | 3, 91, 0, 3, 2, 0, 1 | 1, 6, 1, 5, 0, 0, 1 |
| 10383-fearful_F-AA-15.jpg | 2, 1, 82, 2, 1, 6, 6 | 0, 1, 10, 0, 0, 0, 3 |
| 10461-fearfulopen_F-AA-15.jpg | 2, 3, 58, 2, 3, 1, 31 | 0, 0, 5, 0, 0, 0, 9 |
| 10526-happy_F-AA-15.jpg | 1, 0, 0, 96, 2, 1, 0 | 0, 0, 0, 0, 0, 0, 0 |
| 10739-neutral_F-AA-15.jpg | 1, 0, 1, 89, 7, 1 | 0, 0, 0, 0, 0, 0, 0 |
| 10867-neutralopen_F-AA-15.jpg | 2, 2, 10, 1, 33, 0, 52 | 0, 0, 0, 0, 7, 0, 7 |
| 10967-sad_F-AA-15.jpg | 3, 3, 6, 1, 2, 85, 0 | 2, 0, 0, 0, 0, 0, 0 |
| 11027-sadopen_F-AA-15.jpg | 0, 5, 22, 0, 72, 1 | 0, 0, 0, 0, 0, 0, 0 |
| 11079-surprise_F-AA-15.jpg | 1, 0, 23, 0, 2, 0, 74 | 0, 0, 1, 0, 0, 0, 13 |

**Table 1** CAFE original annotator distribution vs. filtered worker’s distribution for subject F-AA-15 in CAFE

---

| Image | CAFE labeler distribution (count) | Crowdsourced labeler distribution (count) |
|-------|----------------------------------|------------------------------------------|
| 9979-angry_F-AA-03.jpg | 89, 4, 0, 0, 4, 3 | 78, 31, 2, 8, 3, 2, 3 |
| 10100-angryopen_F-AA-03.jpg | 16, 0, 36, 5, 1, 2, 40 | 17, 3, 38, 15, 0, 0, 54 |
| 10184-disgust_F-AA-03.jpg | 17, 41, 2, 12, 19, 8, 1 | 12, 75, 1, 10, 25, 3, 1 |
| 10280-disgustwithtongue_F-AA-03.jpg | 19, 77, 0, 0, 2, 1, 1 | 19, 85, 2, 13, 6, 0, 2 |
| 10375-fearful_F-AA-03.jpg | 2, 4, 49, 13, 2, 3, 27 | 10, 12, 27, 15, 16, 3, 44 |
| 10454-fearfulopen_F-AA-03.jpg | 0, 1, 27, 4, 1, 0, 67 | 1, 0, 43, 3, 0, 0, 80 |
| 10515-happy_F-AA-03.jpg | 1, 0, 98, 1, 0, 0 | 0, 0, 113, 8, 1, 1 |
| 10635-happyopen_F-AA-03.jpg | 1, 2, 0, 94, 1, 0, 2 | 0, 0, 0, 126, 0, 0, 1 |
| 10730-neutral_F-AA-03.jpg | 3, 2, 0, 2, 73, 19, 1 | 2, 1, 0, 77, 47, 0 |
| 10858-neutralopen_F-AA-03.jpg | 4, 4, 4, 2, 17, 3, 66 | 1, 7, 21, 1, 10, 5, 82 |
| 10960-sad_F-AA-03.jpg | 2, 1, 2, 1, 2, 92, 0 | 8, 9, 3, 0, 4, 103, 0 |
| 11021-sadopen_F-AA-03.jpg | 1, 6, 21, 15, 13, 30, 14 | 9, 21, 24, 14, 12, 43, 4 |
| 11068-surprise_F-AA-03.jpg | 2, 1, 13, 20, 1, 0, 63 | 1, 1, 12, 30, 0, 1, 82 |

**Table 2** CAFE original annotator distribution vs. unfiltered crowdsourced class distribution for subject F-AA-93 in CAFE
distances between these two groups is statistically significant according to an independent 2-sample t-test ($t = 3.2827$, $p = 0.0014$). To visualize this difference, Fig. 4 compares the true emotion distribution with the emitted distribution of each of the two classifiers for 3 representative images in the test set with subjective labels.

**Discussion**

Interaction designers and developers of affective computing systems should consider whether soft or hard targets are the most appropriate label representation for training an affective computer vision classifier for a particular application and dataset. Affective computer vision models which are optimized for understanding the potentially diverse range of human interpretations of emotion can be used in several applications of interactive systems, such as AI-powered systems which aid individuals with autism and other developmental delays [25, 26, 73–86], e-learning systems [87, 88], or at-home diagnostic screening tools for psychiatry conditions [89–105].

There are several limitations of this work. This study was performed on a single dataset. For these results to generalize to other types of images, including for crowdsourced soft-target label generation in domains outside of emotion recognition, other datasets must be explored. Another limitation is that we did not record or account for potential biases in the quality control steps for filtering the crowd. Further study into how differing crowd quality mechanisms affect the result would be interesting, as the data label quality can drastically affect a machine learning algorithm. A final limitation is that we did not have a reliable method to disentangle compound emotions from ambiguous labels.

By acquiring labels from 100 independent crowd workers per image, each providing their vote for the most prominent emotion, we created a representation describing the between-subject subjectivity of the emotion expressed in the image. This representation notably obfuscates within-subject subjectivity, and an alternative which should be studied in future work is to ask each rater to provide multiple selections through a semantical scale, as in Korovina et al. [55, 56].
Conclusion

For many applications of affective computing, reporting an emotion probability distribution that accounts for the subjectivity of human interpretation can be more important than traditional machine learning metrics. Crowdsourcing is a feasible solution for acquiring soft-target labels provided a sufficient filtering mechanism for selecting reliable crowd workers.

Funding This work was supported in part by funds to DPW from the National Institutes of Health (1R01EB025025-01, 1R21HD091500-01, 1R01LM013083, 1R01LM013364), the National Science Foundation (Award 2014232), The Hartwell Foundation, Bill and Melinda Gates Foundation, Coulter Foundation, Lucile Packard Foundation, the Weston Havens Foundation, and program grants from Stanford’s Human Centered Artificial Intelligence Program, Stanford’s Precision Health and Integrated Diagnostics Center (PHIND), Stanford’s Beckman Center, Stanford’s Bio-X Center, Predictives and Diagnostics Accelerator (SPADA) Spectrum, Stanford’s Spark Program in Translational Research, Stanford mediaX, and Stanford’s Wu Tsai Neurosciences Institute’s Neuroscience: Translate Program. We also acknowledge generous support from David Orr, Imma Calvo, Bobby Dekesyer and Peter Sullivan. P.W. would like to acknowledge support from Mr. Schroeder and the Stanford Interdisciplinary Graduate Fellowship (SIGF) as the Schroeder Family Goldman Sachs Graduate Fellow.

Declarations

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional (approved by the Stanford University Institutional Review Board) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent was obtained from all individual participants (crowd workers on Microworkers.com) included in the study.

Conflict of Interest Dr. Dennis P. Wall is the scientific founder of Cognoa, a company focused on digital pediatric healthcare; the approach and findings presented in this paper are independent from/not related to Cognoa. All the other authors have declared no competing interests exist.

References

1. Cambria E, Dipankar D, Sivaji B, Antonio F. Affective computing and sentiment analysis. In A practical guide to sentiment analysis, pp. 1–10. Springer, Cham, 2017.
2. Hupont I, Sandra B, Eva C, Rafael DH, Advanced human affect visualization. In 2013 IEEE International Conference on Systems, Man, and Cybernetics, pp. 2700-2705. IEEE, 2013.
3. Jerald R. Wearable emotion detection and feedback system. U.S. Patent 9,019,174, issued April 28, 2015.
4. Liu R, Salisbury JP, Vahabzadeh A, Sahin NT. Feasibility of an autism-focused augmented reality smartglasses system for social communication and behavioral coaching. Front Pediatr. 2017;5:145.
5. Völkel ST, Julia G, Ramona S, Renate H, Clemens S, Quay A, Heinrich H. I drive my car and my states drive me: visualizing driver’s emotional and physical states. In Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 198-203. 2018.
6. Kaur R, Sandeep K. Multimodal sentiment analysis: a survey and comparison.International Journal of Service Science, Management, Engineering, and Technology (IJSSMET). 2019;10(2):38-58.
7. Poria S, Erik C, Alexander G. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the 2015 conference on empirical methods in natural language processing, pp. 2539-2544. 2015.
8. Poria S, Cambria E, Howard N, Huang G-B, Hussain A. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing. 2016;174:50–9.
9. Tahir M, Abdallah T, Feras AO, Babar S, Zahid H, Muhammad W. A novel binary chaotic genetic algorithm for feature selection and its utility in affective computing and healthcare. Neur Comp Appl. 2020:1-22.
10. Yannakakis GN. Enhancing health care via affective computing. 2018.
11. Ebyen F, Martin W, Tony P, Bjorn S, Christoph B, Berthold F, Nhu NT. Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car. Advances in human-computer interaction. 2010.
12. Devillers L, Vidrascu L, Lamel L. Challenges in real-life emotion annotation and machine learning based detection. Neural Netw. 2005;18(4):407–22.
13. Zhang L, Steffen W, Xueyao M, Philipp W, Ayoub AH, Harald CT, Sascha G. BioVis Emo DB: a multimodal database for emotion analyses validated by subjective ratings. In2016 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1-6. IEEE, 2016.
14. Zhou Y, Xuefeng L, Yu G, Yi Fei Y, Longshan Y. Multi-classifier interactive learning for ambiguous speech emotion recognition. 2020. arXiv preprint https://arxiv.org/abs/2012.05429.
15. Magdin M, Prikler F. Real time facial expression recognition using webcam and SDK affectiva. IJMAI15. 2018:7-15.
16. McDuff D, Rana K, Thibaud S, May A, Jeffrey C, Rosalind P. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 881-888. 2013.
17. Ando A, Satoshi K, Hosana K, Ryo M, Yusuke I, Yushi A. Soft-target training with ambiguous emotional utterances for DNN-based speech emotion classification. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. pp. 4964-4968. IEEE.
18. Fang X, Jiancheng Y, Bingbing N. Stochastic label refinery: toward better target label distribution. In 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9115-9121. IEEE, 2021.
19. Yin Da, Liu X, Xiuyu Wu, Chang B. A soft label strategy for target-level sentiment classification. Sentiment and Social Media Analysis: In Proceedings of the Tenth Workshop on Computational Approaches to Subjectivity; 2019. p. 6–15.
20. Turing AM. Computing machinery and intelligence. In Parsing the turing test, pp. 23-65. Springer, Dordrecht, 2009.
21. Zeng Z, Jilin Tu, Liu M, Huang TS, Pianfetti B, Roth D, Levinson S. Audio-visual affect recognition. IEEE Trans Multimedia. 2007;9(2):424–8.
22. Kratzwald B, Ilie S, Kraus M, Feurriegel S, Prendinger H. Deep learning for affective computing: text-based emotion recognition in decision support. Decis Support Syst. 2018:115:24–35.
23. Tao J, Tieniu T. Affective computing: a review. In International Conference on Affective computing and intelligent interaction, pp. 981-995. Springer, Berlin, Heidelberg, 2005.

24. Haber N, Catalin V, Azar F, Terry W, Dennis PW. A practical approach to real-time neutral feature subtraction for facial expression recognition. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-9. IEEE, 2016.

25. Voss C, Jessy S, Jena D, Aaron K, Nick H, Peter W, Qandeel T et al. Effect of wearable digital intervention for improving socialization in children with autism spectrum disorder: a randomized clinical trial. JAMA pediatrics. 2019;173(5):446-454.

26. Voss C, Peter W, Nick H, Aaron K, Jena D, Azar F, Titas D et al. Superpower glass: delivering unobtrusive real-time social cues in wearable systems. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 1218-1226. 2016.

27. Dinculescu A, Andrea B, Carmen S, Livia P, Cristian V, Alexandru M, Nicoară T, Vlad V. Automatic identification of anthropological face landmarks for emotion detection. In 2019 9th International Conference on Recent Advances in Space Technologies (RAST), pp. 585-590. IEEE, 2019.

28. Nguyen BT, Minh HT, Tan VP, Hien DN. An efficient real-time emotion detection using camera and facial landmarks. In 2017 seventh international conference on information science and technology (ICIST), pp. 251-255. IEEE, 2017.

29. Sharma M, Anand SJ, Aamir K. Emotion recognition using facial expression by fusing key points descriptor and texture features. Multi Tools Appl. 2019;78(12):16195-16219.

30. Fan Y, Jacqueline CKL, Victor OKL. Multi-region ensemble convolutional neural network for facial expression recognition. In International Conference on Artificial Neural Networks, pp. 84-94. Springer, Cham, 2018.

31. Washington P, Haik K, Jack K, Arman H, Aaron K, Emilie L, Cathy H et al. Training an emotion detection classifier using frames from a mobile therapeutic game for children with developmental disorders. 2020. arXiv preprint https://arxiv.org/abs/2012.08678.

32. Ekman P. "Are there basic emotions?" 1992:550.

33. Ekman P. Basic emotions. Handbook of cognition and emotion. 1999:98(45–60):16.

34. Du S, Tao Y, Martinez AM. Compound facial expressions of emotion. Proc Natl Acad Sci. 2014;111(15):E1454-62.

35. Zhang X, Wenzhong Li Xu, Chen, and Sanli Lu. Moodexplorer: towards compound emotion detection via smartphone sensing. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. 2018;1(4):1–30.

36. Lotfian R, Carlos B. Over-sampling emotional speech data based on subjective evaluations provided by multiple individuals. IEEE Transactions on Affective Computing (2019).

37. McFarland DJ, Muhammad AP, William AS, Rita ZG, Jonathan RW. Prediction of subjective ratings of emotional pictures by EEG features. J Neu Eng. 2016;14(1):016009.

38. Rizos G, Björn WS. Average Jane, where art thou?—Recent avenues in efficient machine learning under subjectivity uncertainty. In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 42-55. Springer, Cham, 2020.

39. Villon O, Christine L. Toward recognizing individual’s subjective emotion from physiological signals in practical application. In Twentieth IEEE International Symposium on Computer-Based Medical Systems (CBMS’07), pp. 357-362. IEEE, 2007.

40. Mower E, Mataric MJ, Narayanan S. A framework for automatic human emotion classification using emotion profiles. IEEE Transactions on Audio, Speech, and Language Processing. 2010;19(5):1057–70.

41. Mower E, Angeliki M, Chi CL, Abe K, Carlos B, Sungbok L, Shrikanth N. Interpreting ambiguous emotional expressions. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, pp. 1-8. IEEE, 2009.

42. Fujikawa T, Dario B, Takeshi H, Kenji N. Addressing ambiguity of emotion labels through meta-learning. 2019. arXiv preprint https://arxiv.org/abs/1911.02216.

43. ThiC. Classification on soft labels is robust against label noise. In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, pp. 65-73. Springer, Berlin, Heidelberg, 2008.

44. Yang Z, Liu T, Liu J, Wang Li, Zhao S. A novel soft margin loss function for deep discriminative embedding learning. IEEE Access. 2020;8:20785–94.

45. Peterson JC, Ruairidb MH, Thomas LG, Olga R. Human uncertainty makes classification more robust. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9617-9626. 2019.

46. Uma A, Fornaciari T, Hovy D, Paun S, Plank B, Poesio M. A case for soft loss functions. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 2020;8(1):173–7.

47. Chaturvedi V, Satapathy R, Cavallari S, Cambria E. Fuzzy commonsense reasoning for multimodal sentiment analysis. Pattern Recogn Lett. 2019;125:264–70.

48. Nicolaou MA, Gunes H, Pantel M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. IEEE Trans Affect Comput. 2011;2(2):92–105.

49. Parchasaraiah S, Basso C. Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning. In Interspeech. 2017:2013:1103–7.

50. Stappen L, Baird A, Cambria E, Schuller Björn W. Sentiment analysis and topic recognition in video transcriptions. IEEE Intell Syst. 2021;36(2):88–95.

51. Yu LC, Jin W, Robert LK, Xue-jie Z. Predicting valence-arousal ratings of words using a weighted graph method. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 788-793. 2015.

52. Zhao S, Hongxun Y, Xiaolei J. Predicting continuous probability distribution of image emotions in valence-arousal space. In Proceedings of the 23rd ACM international conference on Multimedia, pp. 879-882. 2015.

53. Kairam S, Jeffrey H. Parting crowds: characterizing divergent interpretations in crowdsourced annotation tasks. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, pp. 1637-1648. 2016.

54. Rodrigues F, Francisco P. Deep learning from crowds. In Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1).

55. Korotkova O, Fabio C, Radoslaw N, Marcos B, Olga B. Investigating crowdsourcing as a method to collect emotion labels for images. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 1-6. 2018.

56. Korotkova O, Baez M, Casati F. Reliability of crowdsourcing as a method for collecting emotions labels on pictures. BMC Res Notes. 2019;12(1):1–6.

57. LoBue V, Baker L, Thrasher C. Through the eyes of a child: preschoolers’ identification of emotional expressions from the child affective facial expression (CAFE) set. Cogn Emot. 2018;32(5):1122–30.

58. LoBue V, Thrasher C. The Child Affective Facial Expression (CAFE) set: validity and reliability from untrained adults. Front Psychol. 2015;5:1532.

59. Paolacci G, Chandler J, Ipeirotis PG. Running experiments on amazon mechanical turk. Judgm Decis Mak. 2010;5(5):411–9.
60. Hirth M, Tobias H, Phuoc TG. Anatomy of a crowdsourcing platform using the example of microworkers. In 2011 Fifth international conference on innovative mobile and internet services in ubiquitous computing. pp. 322-329. IEEE, 2011.

61. Coolican J, Eskes GA, McMullen PA, Lecky E. Perceptual biases in processing facial identity and emotion. Brain Cogn. 2008;66(2):176–87.

62. Coren S, Russell JA. The relative dominance of different facial expressions of emotion under conditions of perceptual ambiguity. Cogn Emot. 1992;6(5):339–56.

63. Gray, Katie LH, Wendy JA, Nicholas H, Kristiana E, Newton, and Matthew Garner. Faces and awareness: low-level, not emotional factors determine perceptual dominance. Emotion. 2013;13(3):537.

64. Allabakhsh M, Boualem B, Aleksandar I, Hamid RMM, Elisa B, Schahram D. Quality control in crowdsourcing systems: issues and directions. IEEE Internet Computing 2013;17(2):76-81.

65. Buchholz S, Javier L. Crowdsourcing preference tests, and how to detect cheating. In 12th Annual Conference of the International Speech Communication Association. 2011.

66. Daniel F, Kucherbaev P, Cappiello C, Benatallah B, Allabakhsh M. Quality control in crowdsourcing: a survey of quality attributes, assessment techniques, and assurance actions. ACM Computing Surveys (CSUR). 2018;51(1):1–40.

67. Lease M. On quality control and machine learning in crowdsourcing. Human Computation. 2011;11(11).

68. He K, Xiangyu Z, Shaoqing R, Jian S. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 777–785. 2016.

69. Deng J, Wei D, Richard S, Li-Jia L, Kai L, Li FF.Imagenet: a large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248-255. IEEE, 2009.

70. Chollet F. Keras: the python deep learning library. ascl. ascl-1806. 2018.

71. Abadi M, Paul B, Jianmin C, Zhifeng C, Andy D, Jeffrey D, Matthieu D et al. Tensorflow: a system for large-scale machine learning. In 12th {USENIX} symposium on operating systems design and implementation (OSDI 16). 2016. pp. 265-283.

72. Kingma DP, Jimmy B. Adam: a method for stochastic optimization. 2014. arXiv preprint https://arxiv.org/abs/1412.6980.

73. Daniels J, Schwartz JN, Voss C, Haber N, Fazel A, Kline A, Washington P, Feinstein C, Winograd T, Wall DP. Exploratory study examining the at-home usability of a wearable tool for autism and ADHD. Transl Psychiatry. 2015;5(2):e514–e514.

74. Daniels J, Nick H, Catalin V, Jessey S, Serena T, Azar F, Aaron K et al. Feasibility testing of a wearable behavioral aid for social learning in children with autism. J Med Int Res. 2019;21(4):e13822.

75. Kosmicki JA, Sochat V, Duda M, Wall DP. Searching for a minimal set of behaviors for autism detection through a feature selection-based machine learning. Transl Psychiatry. 2015;5(2):e514–e514.

76. Leblanc E, Washington P, Varma M, Dunlap K, Penev Y, Kline A, Wall DP. Feature replacement methods enable reliable home video analysis for machine learning detection of autism. Sci Rep. 2020;10(1):1-11.

77. Tariq Q, Scott LF, Jessey NS, Kaitlyn D, Connor C, Peter W, Haik K, Naila ZK, Gary LD, Dennis PW. Detecting developmental delay and autism through machine learning models using home videos of Bangladeshi children: development and validation study. J Med Int Res. 2019;21(4):e13822.

78. Tariq Q, Jena D, Jessey NS, Peter W, Haik K, Dennis PW. Mobile detection of autism through machine learning on home video: a development and prospective validation study. PLoS medicine. 2018;15(11):e1002705.
Authors and Affiliations

Peter Washington1 · Haik Kalantarian2 · Jack Kent2 · Arman Husic2 · Aaron Kline2 · Emilie Leblanc2 · Cathy Hou3 · Cezmi Mutlu4 · Kaitlyn Dunlap2 · Yordan Penev2 · Nate Stockham3 · Brianna Chrisman1 · Kelley Paskov6 · Jae-Yoon Jung2 · Catalin Voss3 · Nick Haber7 · Dennis P. Wall8

Haik Kalantarian
haik.kalantarian@gmail.com

Jack Kent
jackkent@stanford.edu

Arman Husic
ahusic@stanford.edu

Aaron Kline
akline@stanford.edu

Emilie Leblanc
emilie.leblanc@stanford.edu

Cathy Hou
cathyhou@stanford.edu

Cezmi Mutlu
cezmi@stanford.edu

Kaitlyn Dunlap
kaiti.dunlap@stanford.edu

Yordan Penev
ypenev@stanford.edu

Nate Stockham
stockham@stanford.edu

Brianna Chrisman
briannac@stanford.edu

Kelley Paskov
kpaskov@stanford.edu

Jae-Yoon Jung
jaeyjung@stanford.edu

Catalin Voss
catalin@cs.stanford.edu

Nick Haber
nhaber@stanford.edu

1 Department of Bioengineering, Stanford University, Stanford, USA
2 Department of Pediatrics (Systems Medicine), Stanford University, Stanford, USA
3 Department of Computer Science, Stanford University, Stanford, USA
4 Department of Electrical Engineering, Stanford University, Stanford, USA
5 Department of Neuroscience, Stanford University, Stanford, USA
6 Department of Biomedical Data Science, Stanford University, Stanford, USA
7 Graduate School of Education, Stanford University, Stanford, USA
8 Departments of Pediatrics (Systems Medicine), Biomedical Data Science, and Psychiatry and Behavioral Sciences, Stanford University, Stanford, USA