Towards Automated Classification of Emotional Facial Expressions

Lewis J. Baker (lewis.j.baker@rutgers.edu)[1], **Vanessa LoBue (vlobue@rutgers.edu)**[2],
Elizabeth Bonawitz (elizabeth.bonawitz@rutgers.edu)[2], & **Patrick Shafto (patrick.shafto@gmail.com)**[1]

[1]Department of Mathematics and Computer Science, [2]Department of Psychology
Rutgers University – Newark, 101 Warren St., Newark, NJ, 07102 USA

Abstract

Emotional state influences nearly every aspect of human cognition. However, coding emotional state is a costly process that relies on proprietary software or the subjective judgments of trained raters, highlighting the need for a reliable, automatic method of recognizing and labeling emotional expression. We demonstrate that machine learning methods can approach near-human levels for categorization of facial expression in naturalistic experiments. Our results show relative success of models on highly controlled stimuli and relative failure on less controlled images, emphasizing the need for real-world data for application to real-world experiments. We then test the potential of combining multiple freely available datasets to broadly categorize faces that vary across age, race, gender and photographic quality.

Keywords: Classification, machine learning, computer vision, support vector machines, emotion and cognition, facial recognition

Introduction

Emotions are widely assumed to play a causal role in nearly every aspect of cognition (e.g., Pessoa, 2008), and yet many studies in cognitive science (and developmental science in particular) neglect to measure emotion because current measures are either expensive, tedious, or inaccurate. Consequently, many standard practices in the field have turned to indirect measures of affect. One prevalent example is the association of infant looking time with vastly different emotions depending on the researchers' theoretical stance, including preference (e.g., to positive emotional expressions; LaBarbera, Izard, Vietze, and Parisi, 1976), interest (e.g., to animate stimuli; Csibra, 2008), or surprise (e.g., to violations of belief; Baillargeon, Scott, and He, 2010). Another example is the notable lack of emotional state measures in studies on attention, learning and memory, even though the field has acknowledged the impact of emotion on these functions for over 50 years (Easterbrook, 1959). Rather than inferring the role of emotions, future studies could measure it efficiently using facial recognition algorithms. The advent of elegant machine-learning algorithms offers a free, reliable, non-invasive and easily implemented method that may be able to measure affective state in real-world settings at levels that meet or exceed trained human raters.

Here, we demonstrate automatic classification of emotional faces using three different datasets. We concentrate on young populations, as developmental science is particularly interested and constrained by hand-coding, but also demonstrate that methods are easily extended to adult populations. We also concentrate on relatively simple machine learning algorithms that may be flexibly implemented for a variety of psychological studies. In doing so, we highlight the need for real-world data to solve real-world problems, as models based on well-curated training images that are common in the field often fail to accurately categorize messy, uncontrolled images. We further show how a single, large dataset that leverages controlled and uncontrolled images can improve generalization to real-world stimuli.

Recognition of facial expressions is a useful, non-invasive method of reasoning about another's thoughts. The seeming universality of emotional expressions further underscores their importance (Ekman & Friesen, 1971). However, to understand how emotion influences cognition, researchers must be able to categorize facial expressions in continuous time – and no existing measure can do this without great expense of time or money. Participant surveys lack temporal resolution and fall prey to metacognitive errors. Physiological methods require expensive or invasive apparatus such as galvanic skin response monitors or cortisol measurements Picard, Vyzas, and Healey, 2001. Even the gold-standard method of the Facial Action Coding System (FACS; Ekman and Rosenberg, 1997) requires hours of effort by trained technicians or prohibitively expensive proprietary software. As the demand for ecological experimentation increases, so too does the volume of video data for researchers (or more often their students) to scrutinize and label, frame by frame. The relatively constrained problems of identifying, labeling and categorizing facial features over thousands of datapoints is a prime opportunity for a machine learning solution.

Advances in data science and machine learning offer an affordable and accurate measure of participant emotional state using only filmed recordings. Computer scientists have demonstrated the uncanny accuracy of basic algorithms to classify highly controlled emotional images (e.g., Cohn, Zlochower, Lien, and Kanade, 1999; Littlewort, Bartlett, Fasel, Susskind, and Movellan, 2006), and recent efforts to categorize emotion "in the wild" (Yao, Shao, Ma, & Chen, 2015) have the problem to unsupervised learning for less controlled images. Here we are interested in applying machine-learning to the varied contexts typical of cognitive science experiments. Laboratory settings offer more control than streaming surveillance footage, but less control than posed photography. We approximate this by comparing human performance to an algorithm trained on three datasets with unique attributes, each of which could reasonably be applied to experimental settings. We demonstrate the need for large amounts of highly varied data to consistently and accurately categorize human facial expressions. Furthermore, we present a model

trained on images that vary by age, ethnicity, gender and photographic conditions that nonetheless approaches human rater performance. It is worth noting that our goal is not to model human performance or develop new machine-learning methods; rather, we wish to explore the kinds of data required to approximate human-level emotion coding for cognitive experiments.

We begin by introducing several datasets with unique attributes of in interest to different applications. We define the methods required to use open-source libraries to create a simple machine-learning classifier. Applying this model to the datasets reveals that highly controlled training stimuli are more easily categorized, and that noisier, real-world stimuli are unsurprisingly more difficult. We discuss the trade-off between accuracy and generality by amalgamating three datasets into a comprehensive model that is more robust to noisy input.

General Methods

Databases

Machine learning requires a large number of samples for reliable classification. However, the type of input can greatly affect the generalizability of the model. For instance, a model trained on only children's faces might not perform well with adult faces. Likewise, a model trained on highly controlled images might not perform well on naturalistic stimuli. We drew from three sources for training, each with a particular strength that could improve performance in a given setting. A "face" was defined as a front-facing image containing two eyes and no obstructions to facial features.

The CAFE dataset. Most face databases for psychology and machine learning focus on adults. However, a recent effort by LoBue and Thrasher (2015) documented the facial expressions of young children for applications in developmental psychology. Although stimulus sets exist for older children (aged 8-17, Egger et al., 2011) and adults (Cohn et al., 1999), the Child Affective Facial Expression (CAFE) set is the only collection featuring young children. The set contains photographs of 154 racially and ethnically diverse 2- to 8-year-old children posing for six emotional facial expressions (angry, disgusted, fearful, happy, sad, and surprised) as well as a resting neutral expression. Facial expressions were further labeled for "open" or "closed" mouths for angry, fearful, happy, sad and neutral faces. Disgust expressions were uniquely coded as with or without a protruding tongue. The CAFE set features multiple emotional faces for each child, though not every child demonstrated every subcategory of emotion. Altogether, the set contains 1192 images. Children's facial features offer a great deal of variability, and the ethnic diversity of the participant sample approximates the demographics of the United States.

The CAFE set was validated by a group of 100 independent adult raters, who viewed each image and labeled it with one of the seven emotions. Importantly, the images are labeled by the expression the child was asked to give, and not by the labels most often generated by the raters. This variability makes the CAFE useful to compare to computer models, as we can test the model's success on "difficult" or "easy" faces compared to human performance.

The CK+ dataset. One method of producing cleaner data for machine learning is to extract images with tightly controlled visual features. Although our focus is on developmental populations, the CAFE set is the only publicly available database of children's faces. We therefore included a dataset comprised only of extensively vetted faces: the Cohn-Kanade AU-Coded Expression Database, Version 2 (Lucey et al., 2010). The Cohn-Kande dataset (CK+) consists of over 11,000 image sequences of 120 adult models as they changed from neutral resting faces to peak emotional expression from 7 categories (the same as the CAFE expressions, with the addition of contempt). It is currently unknown whether training images from an adult dataset would improve performance on child facial categorization. Given the abundance of adult datasets, any improvement on child facial classification would expand the available training data for future models.

Machine vision researchers often use the CK+ dataset as a benchmark for performance of an algorithm (e.g., Littlewort et al., 2006). For our purposes, training the algorithm on the CK+ dataset allows us to test the best case scenario of facial classification, as it contains only highly controlled images with little cross-category variability. This comes at a cost to ecological validity, as all faces are of adults aged 18 to 30, and less than 18% were minorities. Additionally, whereas items in the CAFE set were validated using subjective judgments from adult raters, the peak faces from the CK+ database were validated using the Facial Action Coding System (FACS). Briefly, FACS categorizes faces into emotional categories using reliable expressions of specific facial motor groups, or action units (Ekman & Rosenberg, 1997). Lucey et al. (2010) validated the emotional labels given to each peak face using a linear support vector classifier trained on action units. Selecting these initial and peak faces generated 308 emotional expressions from the 6 emotional categories with a corresponding neutral face for each. A single neutral face was randomly selected for each participant to prevent over-fitting of neutral faces, leaving 120 neutral faces and a total of 428 faces.

Google image search by category. The CAFE and CK+ sets feature images taken under ideal lighting and camera positions, with labels that have been rigorously validated. However, real-world use of a facial expression classifier would necessarily include less-than-ideal photographic circumstances. To approximate the noisiness of real-world stimuli, we extracted images from a Google image search with the search term "X child face", where X was an emotional category of interest. Images were selected by research assistants, with the criteria that each image featured an individual human child's face (approximately aged 3-10) without obstruction on the face area.

Research assistants terminated the collection of images if the total number of collected images exceeded 100 exemplars or if the search returned more than 20 images in a row without a viable exemplar. This produced only 2 neutral exemplars, so an additional search was conducted using "calm" and "serious" as additional terms for neutral. This produced a total of 609 faces from all seven categories

Face Extraction

Images from all datasets contained extraneous information, including body parts (e.g., hair and shoulders) or photographic artifacts (e.g., serial numbers in the CAFE and CK+ datasets, miscellaneous objects in the Google dataset). All images were passed through a facial recognition algorithm [1] and reduced to a 300 x 300 pixel rectangle centered on the identified face.

Facial recognition was conducted using Haar Feature-based Cascade Classifiers. Generally, the cascade classifier breaks an image into clusters of pixels and excludes clusters that do not resemble facial features from later analysis. The process is then repeated until only clusters that resemble facial features remain. For methodological details and validation, see Viola and Jones (2001). The end result is a computationally efficient method for identifying facial regions.

A trained cascade classifier was obtained from the OpenCV website (Itseez, 2016). All faces from all datasets were passed through the classifier and cropped. A member of the research team then examined each extracted face and discarded false positives on non-face objects. This method produced 1187 faces (5 removed) from the CAFE set, 427 faces (1 removed) from the CK+ dataset, and 477 faces (132 removed) from the Google dataset.

Human Validation

Image category labels for the CAFE and CK+ datasets were validated using adult human raters. To ensure that all images were of equal quality when training the classifier, we validated the Google dataset using 87 adult human raters recruited via Amazon Mechanical Turk. This was necessary to compare classifier and human performance for images that more closely resemble the real world.

Raters (median age: 31; 61 females, 41 college graduates, 28 parents) labeled a representative subset of the Google faces (between 47 and 49 images, evenly distributed across categories) into one of seven emotional categories. Raters also labeled a subset of the CAFE dataset (42 images, 6 from each category). The CAFE faces were evenly distributed by difficulty according to the CAFE set's previous validation metrics. This was done to compare the performance of in-person (live) and online raters.

[1]The facial recognition algorithm was adapted from the Open Source Computer Vision Library (OpenCV v2.4.13; Bradski, 2000) and programmed in Python 2.7. OpenCV is an open source library that provides a common infrastructure to machine vision applications in academia and industry.

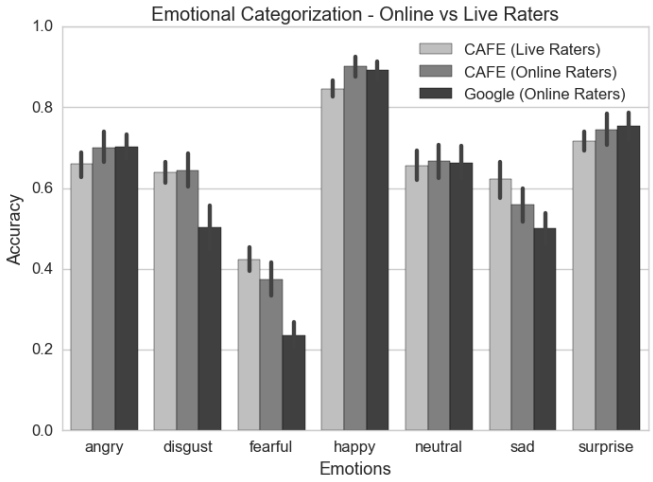

Figure 1: Validation of the CAFE and Google datasets by "online" human raters via Mechanical Turk vs in-person, "live" raters.

We first confirmed that Mechanical Turk raters performed comparably to live human raters (Figure 1). Overall accuracy of ratings for CAFE set images between live and online raters was significantly positively correlated ($r = .783, t_{412} = 25.53, p < .0001$). Furthermore, accuracy of online and live raters were further correlated for all emotional categories (lowest $r_{surprise} = .410, t_{55} = 3.340, p = .001$; highest $r_{sad} = .820, t_{57} = 10.807, p < .0001$), with the exception of happy expressions ($r = .22, t_{57} = 1.742, p = .08$), which likely had a reduced correlation due to ceiling effects. These results support the use of Mechanical Turk raters to validate the Google dataset.

We then confirmed that the human categorization performance for the novel Google images were comparable to the CAFE images. A two-way ANOVA modeling mean online rater categorization performance by dataset and emotional label found significant differences in categorization accuracy by dataset ($F_{1,877} = 8.753, p = .003, \eta^2 = .086$) and emotion ($F_{6,877} = 89.504, p < .0001, \eta^2 = .881$). There was also a significant interaction ($F_{6,877} = 2.366, p = .028, \eta^2 = .023$), indicating no significant difference between online and live raters for angry, happy, neutral, sad and surprised expressions (highest t_{877} [sad] $= .9694, p = .167$), and significantly poorer performance by online raters for disgusted and fearful expressions (lowest t_{877} [disgust] $= 2.134, p < .017, d = .144$). These results suggest that the Google dataset is comparable to the CAFE set for five of seven emotions and follows the same trends for sadness and disgust, making the Google set ideal for testing an algorithm on ecological images.

The Machine Learning Algorithm

Whole research communities are dedicated to the application of machine learning to emotional recognition, using both supervised and unsupervised algorithms for still image, video, audio or multimedia data (e.g., the EmotiW challenge at the annual ACM ICMI conference). Although there have been re-

cent successes modeling dynamic features (Littlewort et al., 2006), We opted to analyze still images to simplify implementation for the target demographic of psychologists, and used a supervised learning approach due to the relatively few training images available. We therefore selected a Support Vector Machine (SVM) algorithm, as SVMs are ideal for use by non-computer scientists for their simple implementation and ease of interpretation. SVMs have a long and successful history in image recognition (Tong & Chang, 2001), particularly with facial recognition (Osuna, Freund, & Girosit, 1997). An SVM is a type of supervised learning in which the algorithm identifies the optimal boundary between labeled data points. The boundary is defined by the support vectors, the subset of the data that define the boundary between classes. This boundary can then be used to infer, based on observed features, which category a novel image (or novel images) best fit. We recommend Cristianini and Shawe-Taylor (2000) for an in-depth overview.

We trained an SVM for each dataset, as well as on a comprehensive dataset containing training images from all three datasets. We used an open-source SVM classifier available through the scikit-learn database (Pedregosa et al., 2011). SVMs require the user to choose a similarity function, called a kernel, that governs the complexity of the possible boundaries between classes. There are many standard options for kernels including linear, polynomial, and radial basis function (RBF; aka Gaussian). Each computes similarity somewhat differently and they consequently differ in the kinds of classification boundaries they admit; as one might expect, a linear kernel gives a linear boundary and polynomial and RBF kernels allow non-linear boundaries. While these non-linear methods offer increased expressiveness, they also increase the risk of overfitting.

Additionally, there are two parameters that must be set and affect outcomes: the regularization parameter C and kernel coefficient γ. C is a regularization parameter which, when set to higher values, allows more complex solutions. The kernel coefficients γ affect the influence of specific specific supports. When γ is small, a support has broad influence on classification decisions, whereas when γ is large the influence of each support is localized to the area near the supporting data point. A grid search for kernels, {linear, polynomial, radial basis function (RBF)}, penalty parameters, $C = (.001, .01, .1, 1, 10, 100)$, and kernel coefficients, $\gamma = (.0001, .001, .01, .1, 1, 10, 100, 1000)$, yielded the optimal combination of a polynomial kernel with a $C = 1$ and a $\gamma = .0001$, as assessed via cross-validation.

Although SVM classifiers are often used for facial recognition, training a classifier for emotional features offers unique problems. The classifier might divide faces by other similarities; emotional expressions are but a subset of the considerable variability between faces. For example, say a classifier is trained on two stimuli: Child A with an angry expression and Child B with a happy expression. When presented with a test image of Child A making a happy expression, the SVM may be more likely to categorize the test image by the stable facial similarities of Child A than to the desired similarities in emotional expression of Child B. One solution might be to randomly select only one face per child participant in the CAFE and CK+ sets. This is not ideal, as it would greatly reduce the training set. Instead, we trained the SVM on all faces for a proportion of participants and tested on all faces for the withheld subset of participants. This eliminated the possibility that a test image might be paired with a training image of the same child, while also maximizing the richness of the dataset.

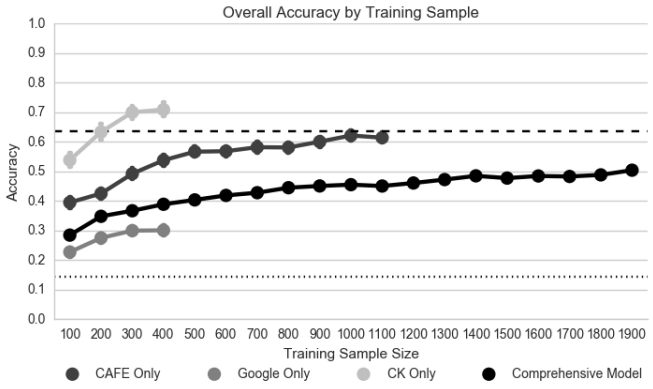

Figure 2: Classification of test images improved as a function of training data. The top line denotes average human accuracy across live and online human validation; the lower line denotes chance.

Another issue unique to emotional classification is the breadth of expression. For instance, the CAFE set makes a distinction between faces with open and closed mouths, and both the CAFE and Google sets contain exemplars that were difficult to label by human raters. We opted to include all instances under the basic emotional category, regardless of subordinate labels or validation score, so as to maximize training data with the greatest possible variation between features.

Results

The algorithm was trained on incrementally increasing sizes of training data from all three datasets individually and a comprehensive dataset trained from all sources. Each sample size by dataset was repeated 40 times with a new random selection of training and test data to approximate error.

The primary goal of these analyses was to demonstrate machine-learning categorization on different training data versus human raters. Figure 2 illustrates overall performance by sample size for each of the datasets. An ANCOVA modeling dataset by training sample size revealed a significant effect of dataset ($F_{3,1512} = 1412.39, p < .0001, \eta^2 = .540$) and training sample size ($F_{1,1512} = 1103.76, p < .0001, \eta^2 = .422$), with a significant interaction ($F_{3,1512} = 96.91, p = .0001, \eta^2 = .037$). Paired comparisons revealed that performance on within-dataset models increased faster than the comprehensive dataset as a function of training size (CAFE: $t_{1512} = 12.246, p < .0001, d = .629$; Google:

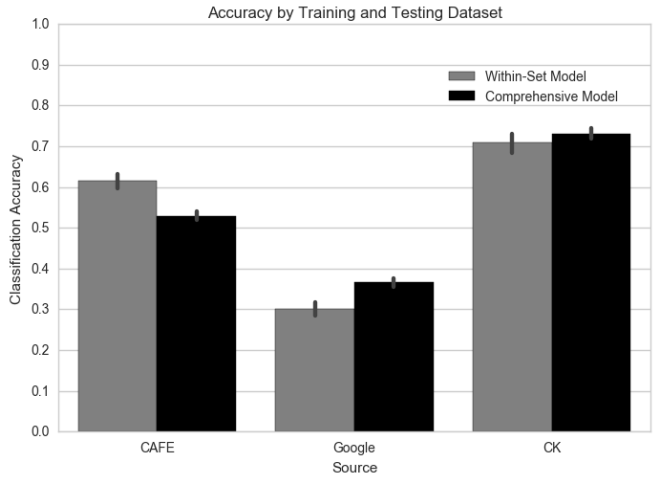

Figure 3: Classification by source of testing data. Accuracy on uncurated Google images improved with the comprehensive model.

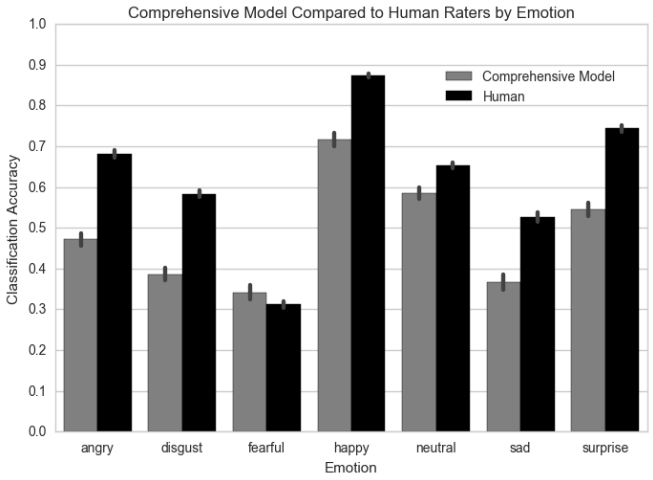

Figure 4: The comprehensive model from all three datasets paralleled human performance.

$t_{1512} = 3.705, p < .0001, d = .191$; CK+: $t_{1512} = 11.886, p < .0001, d = .611$). Altogether, performance on all datasets improved as a function of training data, but performance on within-dataset models increased faster than the comprehensive model.

The bold dotted line on Figure 2 denotes average human categorization performance for the CAFE and Google sets, although it should be noted that no item-wise validation metrics were available for the CK+ set. T-tests revealed that maximum training sizes on the CK+ dataset exceeded human performance ($t_{40} = 6.00, p < .0001, d = 1.90$). Classifier performance was significantly below human performance for the CAFE ($t_{40} = -3.62, p = .0006, d = 1.145$), Google ($t_{40} = -38.65, p < .0001, d = 9.92$), and comprehensive datasets ($t_{40} = -39.51, p < .0001, d = 10.68$). Interestingly, human performance was not significantly correlated with classifier performance at maximum training sample size for any dataset (CK+: $r = -.031, p = .837$; CAFE: $r = .260, p = .105$; Google: $r = -.030, p = .869$; Comprehensive: $r = .240, p = .135$), suggesting that the basis on which categorization decisions were made by the algorithm differed from human judgments.

It is crucial for future applications that a classifier not only categorizes within a training dataset, but can also generalize beyond that set. A common method of gauging generalizability is to train models for each dataset and test on the other datasets. However, all of the present datasets, particularly the CK+ dataset, have unequal numbers of exemplars for each emotional category. As classifier performance is directly related to the amount of training data, we would have to hold training data constant to the minimum possible value across all emotional categories and datasets, which in this case would be only 25 exemplars per category (the number of exemplars for "fear" in the CK+ dataset), for a training set of only 175 images. Instead, we tested how well a single comprehensive model performs against maximally trained models for each individual dataset (the within-set models). This comparison demonstrates how the addition of training images outside the dataset improves performance. An ANOVA comparing accuracy by model type (within-dataset or comprehensive) and source of test images revealed a no effect of model ($F_{1,234} = 0.001, p = .973, \eta^2 < .001$) but a significant effect of test image source ($F_{2,234} = 1175.71, p < .0001, \eta^2 = .961$) as well as a significant interaction ($F_{3,234} = 47.13, p < .0001, \eta^2 = .039$). Accuracy for the comprehensive model was significantly greater than the within-set model for Google ($t_{234} = 6.03, p < .0001, d = .792$), not significantly different for CK+ test images ($t_{234} = 1.09, p = .140, d = .142$) and significantly less for CAFE test images ($t_{234} = 6.50, p < .0001, d = .849$). Comparing Figure 2 and Figure 3, the overall performance deficit of the comprehensive model relative to the CK+ and CAFE sets in Figure 2 are due to the high proportion of training images in the comprehensive model that come from the CAFE set (57.9%). Importantly, these results show that a comprehensive dataset from multiple curated sources improves classification of more realistic and uncontrolled Google set.

Finally, it is worthwhile to see how a comprehensive dataset compares to human raters. Figure 4 compares human and comprehensive model performance by emotional category. An ANOVA modeling emotion by rating type (human vs the comprehensive algorithm) revealed than human raters were significantly more accurate than the classifier ($F_{1,546} = 1450.24, p < .0001, \eta^2 = .567$). There were significant differences by emotion ($F_{6,546} = 1021.81, p < .0001, \eta^2 = .400$) as well as a significant interaction ($F_{6,546} = 84.51, p < .0001, \eta^2 = .048$).

Overall, these results suggests that the comprehensive dataset follows similar trends as human raters. Sampling from multiple datasets improves performance on the highly uncontrolled Google stimuli and the highly controlled CK+ stimuli; although performance drops for the CAFE stimuli, it

likely stems from the high proportion of images in the comprehensive dataset that are sampled from the CAFE. These results suggest that models comprised of more training data may approach human performance on varied images.

Discussion

Psychological theories emphasize the causal role of emotions across a variety of phenomena including learning, memory, and attention. However, emotion is rarely measured in such studies due to the cost, inefficiency and tediousness of modern methods. Widely available and accessible methods for coding emotion would greatly reduce barriers to advancing theory by allowing dense measurement of emotion in continuous time. Using a standard machine learning method, we explored the types of training data one would need to approach human-level coding of the big 7 emotional categories. These included curated image sets developed by psychological researchers and uncontrolled images drawn from Google with crowdsourced labels. We find that comprehensive models generated from multiple datasets improve classification of uncurated images. Overall model performance follows the same trends as human performance, and the inclusion of additional datasets promises to further approach human accuracy.

Cognitive science, and developmental science in particular, are greatly limited by the methods of the day. A typical developmental experiment takes place with one child and one experimenter for fifteen minutes. Such tight controls have led to important insights at the cost of ecological validity. The past 20 years have seen incredible improvements to computational theory and processing power that permit a more flexible study of human behavior. With machine-learning methods, scientists are no longer bound to brief interventions or constrained to discrete conditions. Rather, we can now continuously monitor affect and behavior as a response to the real world. Instead of inferring surprise from an infant's looking times, these models provide a method to measure a reliable indicator of emotion. Instead of assuming a role of affect in student outcomes, we can incorporate emotional expression with an intervention in real time.

This paper represents an effort toward integrating computational methods with cognitive science with the goal of actively measuring all features that support cognition. For now, we have demonstrated the feasibility of using publicly available software and data to code images in minutes rather than days. We have not yet reached human-level performance, but we have shown that the curated datasets that have traditionally been collected improve performance over training on naturalistic uncontrolled images. This marks the first step towards building theories that explain how emotion interacts with cognition in real-world learning scenarios.

Acknowledgments

This research was supported in part by NSF grant CISE-1623486 to L.B., V.L., and P.S.

References

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in cognitive sciences*, *14*(3), 110–118.

Bradski, G. (2000). The open source computer vision library. *Dr. Dobb's Journal of Software Tools*.

Cohn, J. F., Zlochower, A. J., Lien, J., & Kanade, T. (1999). Automated face analysis by feature point tracking has high concurrent validity with manual facs coding. *Psychophysiology*, *36*(1), 35–43.

Cristianini, N. & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition*, *107*(2), 705–717.

Easterbrook, J. A. (1959). The effect of emotion on cue utilization and the organization of behavior. *Psychological Review*, *66*(3), 183.

Egger, H. L., Pine, D. S., Nelson, E., Leibenluft, E., Ernst, M., Towbin, K. E., & Angold, A. (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.

Ekman, P. & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Ekman, P. & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.

Itseez. (2016). Open Source Computer Vision Library. https : / / github.com/itseez/opencv.

LaBarbera, J., Izard, C., Vietze, P., & Parisi, S. (1976). Four-and six-month-old infants' visual responses to joy, anger, and neutral expressions. *Child Development*, *47*(2), 535–538.

Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., & Movellan, J. (2006). Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, *24*(6), 615–625.

LoBue, V. & Thrasher, C. (2015). The child affective facial expression (CAFE) set: Validity and reliability from untrained adults. *Frontiers in Psychology*, *5*, 1532.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE conference on computer vision and pattern recognition* (pp. 94–101).

Osuna, E., Freund, R., & Girosit, F. (1997). Training support vector machines: An application to face detection. In *IEEE computer society conference on computer vision and pattern recognition* (pp. 130–136).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience*, *9*(2), 148–158.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence*, *23*(10), 1175–1191.

Tong, S. & Chang, E. (2001). Support vector machine active learning for image retrieval. In *The 9th ACM international conference on multimedia* (pp. 107–118).

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Vol. 1).

Yao, A., Shao, J., Ma, N., & Chen, Y. (2015). Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 acm on international conference on multimodal interaction* (pp. 451–458). ACM.