Democratizing Psychological Insights from Analysis of Nonverbal Behavior

Daniel McDuff
Microsoft Research
Redmond, USA
damcduff@microsoft.com

Jeffrey M. Girard
Carnegie Mellon University
Pittsburgh, USA
jmgirard@cmu.edu

***Abstract*—The affective computing community has invested heavily in building automated tools for the analysis of facial behavior and the expression of emotion. These tools present a valuable, but largely untapped, opportunity for social scientists to perform observational analyses of nonverbal behavior at very large scale. Various tech companies are collecting huge corpora of images and videos from around the world that could be used to study important scientific questions. However, privacy restrictions and intellectual property concerns render these data inaccessible to most academics. Unfortunately, this limits the potential for scientific advancement and leads to the consolidation of data and opportunity into the hands of a few powerful institutions. In this paper, we ask whether similar psychological insights can be gained by analyzing smaller, public datasets that are more within reach for academic researchers. As a proof-of-concept for this idea, we gather, analyze, and release a corpus of public images and metadata and use it to replicate recent psychological findings about smiling, gender, and culture. In so doing, we provide evidence that psychological insights can indeed by democratized through the automated analysis of nonverbal behavior.**

***Index Terms*—Psychology, smiling, gender, culture, display rules, nonverbal behavior, FACS, Bayesian data analysis**

I. Introduction

Automated analysis of nonverbal behavior holds considerable promise for advancing research in psychology and other social sciences. Using machines to collect and analyze behavioral data allows for far greater efficiency and scalability than traditional methods. Researchers in the past were limited by the amount of data that could be collected and analyzed in labs by research assistants, and observational studies were typically limited to samples of dozens or hundreds of participants. However, technological advances now make it possible to collect and analyze data from thousands or even millions of participants. It is conceivable that large corporations with extensive cloud-computing resources could gather and analyze all the videos on YouTube or all the images on Facebook.

Tech companies are already collecting huge corpora of behavioral data from around the world. For example, Affectiva has compiled several million videos of participants in global market-research settings [1]. However, privacy restrictions and intellectual property concerns often prevent such proprietary datasets from being shared with the scientific community. The data then become a highly valuable but generally inaccessible resource to academic researchers. When data and insights are concentrated within a few companies it tends to create a gulf in power between the data "haves" and the data "have nots."

In order to accelerate research on nonverbal behavior and democratize access to the insights that such research might yield, we propose that publicly available data on the internet, which can be shared openly and analyzed collaboratively, may also be used to study psychological questions.

From a scientific perspective, there are two main challenges for our proposal: (1) How representative-versus-biased is the public behavioral data that can be found online? Can such data yield generalizable insights into human psychology? (2) How valid-versus-noisy are the behavioral measures that can be produced automatically, at scale, and for free? Are the measurements of nonverbal behavior produced by open-source behavior analysis software trustworthy?

We provide an initial exploration of these questions by attempting to replicate previous psychological findings about gender, culture, and smiling in a public dataset collected and analyzed with our proposed method. We chose this topic for our initial exploration for three reasons: it is of substantive interest, we thought these three variables would be relatively easy to measure in online data, and we expected the effects of gender and culture to be rather diffuse and thus protected somewhat from the influence of sampling bias.

A. Smiling, Gender, and Culture

Smiling is a fascinating nonverbal behavior that is both common and nuanced; smiles are typically interpreted as expressing positive emotion, but can actually communicate a wide variety of affective and interpersonal states [2], [3]. There is great interest in developing algorithms to detect and analyze smiles in order to infer these underlying states.

However, research has found that a variety of demographic, cultural, and contextual factors influence when and how a person smiles [1], [4]–[9]. These differences in behavior are thought to be caused by differences in *display rules* that dictate how specific individuals (e.g., women) ought to behave in specific circumstances (e.g., when around children) [10]–[12]. Proper understanding of an individual's behavior thus requires

This material is based upon work partially supported by the National Science Foundation (1722822, 1734868) and National Institutes of Health (MH096951). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of National Science Foundation or National Institutes of Health, and no official endorsement should be inferred.

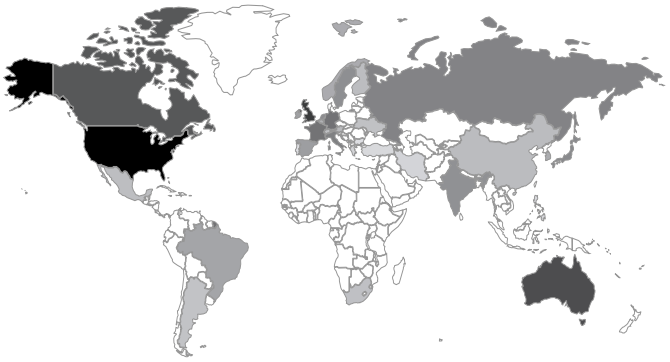

Fig. 1. We analyzed smiling behavior in 290 547 images of 29 579 people from 34 countries. The map shows the countries and distribution of samples used in our study, with darker colors representing more images.

careful consideration of who and where they are, as well as what display rules may be influencing them.

Much of the foundational work on display rules used self-report measures to survey participants about how they typically behave or how they feel they should behave in different circumstances [9], [13]. This work has yielded many interesting findings; in the current study, we focus on two variables that have been strong predictors in recent research. First is an individual's gender: there is considerable evidence that women across the world tend to smile more than men [7], [14]. Second is a culture's history of migration: there is mounting evidence that cultures with more diverse histories of long-term migration (i.e., that have had more source countries contribute to their population over the last 500 years) tend to smile more than cultures with more uniform histories [9]. Cultures higher in this *historical heterogeneity* variable may favor display rules that increase smiling because they involve more interactions between ethnic groups, where higher expressivity may increase trust and help compensate for language barriers [15].

Observational studies that examine the actual behavior of participants from different genders and cultures are an important supplement to self-report studies, but most have historically been limited to rather small samples with dozens or at most hundreds of participants [5], [16], [17].

In the last few years, advances in automated coding and novel approaches to "crowdsourcing" behavioral data have allowed observational studies of cross-cultural differences in facial behavior to increase dramatically in scale. Two recent studies examined video recordings of hundreds of thousands of participants engaged in market research tasks [1], [18]. These studies replicated previous self-report results, finding evidence of many of the expected relationships between smiling, gender, and culture. The authors of these studies described them as exciting advances for behavioral science, but reported that collecting this dataset required collaboration with corporations, took over five years, and cost hundreds of thousands of US dollars [19]. Such an extensive collection of novel behavioral data is not logistically feasible for most researchers, making these advances difficult for other researchers to reproduce.

TABLE I
SUMMARY OF THE DATASET USED IN OUR STATISTICAL ANALYSES

Country	Images	Subjects	Women*	Smile†	HH‡
Argentina	360	114	0.21	0.63	37
Australia	7924	942	0.36	0.98	46
Austria	154	117	0.26	0.93	7
Belgium	426	161	0.22	0.69	10
Brazil	430	214	0.25	0.89	25
Canada	13 949	1433	0.35	0.79	63
China	473	126	0.44	0.85	1
Finland	132	92	0.28	0.65	1
France	4173	886	0.33	0.80	11
Germany	1639	768	0.30	0.76	7
Greece	81	62	0.27	0.56	1
Hong Kong	462	90	0.31	0.84	3
India	1075	388	0.40	0.85	3
Iran	100	61	0.23	0.66	3
Ireland	2733	252	0.24	0.68	12
Israel	619	145	0.31	0.74	22
Italy	1611	562	0.30	0.71	5
Japan	671	403	0.37	0.70	1
Mexico	1027	134	0.40	0.80	25
Netherlands	672	384	0.24	0.83	28
Norway	255	159	0.31	0.77	1
Portugal	91	51	0.14	0.68	15
Puerto Rico	569	53	0.42	0.91	33
Romania	118	69	0.39	0.56	10
Russia	673	448	0.28	0.71	5
South Africa	742	91	0.34	0.74	8
South Korea	486	214	0.47	0.81	1
Spain	1439	290	0.26	0.64	12
Sweden	992	369	0.46	0.89	8
Switzerland	187	88	0.27	1.03	12
Turkey	112	67	0.33	0.47	6
Ukraine	115	94	0.30	0.69	4
UK	28 816	3434	0.31	0.76	25
USA	217 241	16 818	0.34	0.93	83
Total	290 547	29 579			

* Fraction Women † Mean Smile Intensity ‡ Historical Heterogeneity

B. The Current Study

Our goal in the current study was to see if we could replicate the findings of the aforementioned self-report and observational studies using public images scraped from the internet. Specifically, we accessed a public corpus of images of celebrities from around the world and assessed the degree to which these images would reveal evidence of the hypothesized gender and culture display rules (i.e., higher smiling in women and in individuals from countries with higher historical heterogeneity). We reasoned that, if it were possible to recover the same signal in this messy and imperfect sample of behavioral data, it would open up a potentially fruitful avenue for future research that would be far more accessible to most researchers.

Furthermore, we release the meta-data we collected to accompany the public image dataset from which our results were derived, including gender, nationality, and smile intensity labels for 290 547 images of 29 579 celebrities from 34 countries (Figure 1). We hope that the release of these data will encourage researchers to think not only about how public datasets can accelerate algorithmic advances in behavioral coding, but also psychological understanding more broadly.

II. Data

We analyzed a subset of the IMDB-WIKI dataset [20]. This dataset includes over half a million images of international celebrities: 461 871 images of the 100 000 most popular profiles on IMDB and all 62 359 profile images on Wikipedia. Gender labels for most of these celebrities were included in the original dataset; however, nationality labels were not. Because this information was critical to our questions about culture, we attempted to lookup the nationality label for each celebrity in the dataset. We searched the Google Knowledge database for the name of each celebrity and analyzed the queried biographical data using NLTK [21] and named-entity analysis. We were able to extract nationality data for 97.5 % of the celebrities with this approach. After filtering for celebrities for which we had both nationality and gender labels, we were left with a subset of 507 737 images of 70 086 unique individuals. We then dropped all images in which a face was not detected by our smile measurement system (described below) and selected all nationalities (i.e., countries) for which we had usable images from at least 50 different individuals. This selection criteria resulted in our final sample of 290 547 images of 29 579 individuals from 34 countries (Table I).

As in previous work [1], [9], we used the World Migration Matrix (WMM) [22] to calculate the historical heterogeneity of long-term migration cultural variable as the number of source countries that have contributed to each country's present-day population since A.D. 1500. For statistical analysis, this variable was log-transformed and then globally centered by subtracting the mean of all countries in the WMM.

III. Smile Measurement and Validation

The images were analyzed with OpenFace [23], which uses computer vision and machine learning techniques to estimate the intensity of smiles and other facial actions described in the Facial Action Coding System (FACS)[1] [24]. OpenFace uses a Conditional Local Neural Fields (CLNF) model to locate 68 facial landmark points on the face. A similarity transform is used to register the face to a frontal position. The face is then resized to a 112×112 px image with an inter-pupilary distance of 45 px. Histogram of Oriented Gradient (HOG) features are extracted using blocks of 2×2 cells, each 8×8 px, leading to 12×12 blocks of 31-dimensional histograms. To estimate the intensity of facial actions, linear support vector regression models analyze these features. For more details, see [23].

Images scraped from the Internet often contain extreme head pose and lighting conditions, occlusions, and other appearance-related challenges. Thus, we felt it was important to validate the smile intensity estimates on our specific dataset, which may differ in various ways from the datasets that OpenFace was trained and initially validated on. To do so, we selected a subset of 273 images to use for validation and made sure to include a relatively balanced number of images from each country, gender, and estimated smile intensity level.

[1]In FACS, a smile is defined as a contraction of the *zygomaticus major* muscle and is referred to as action unit 12 (AU12).

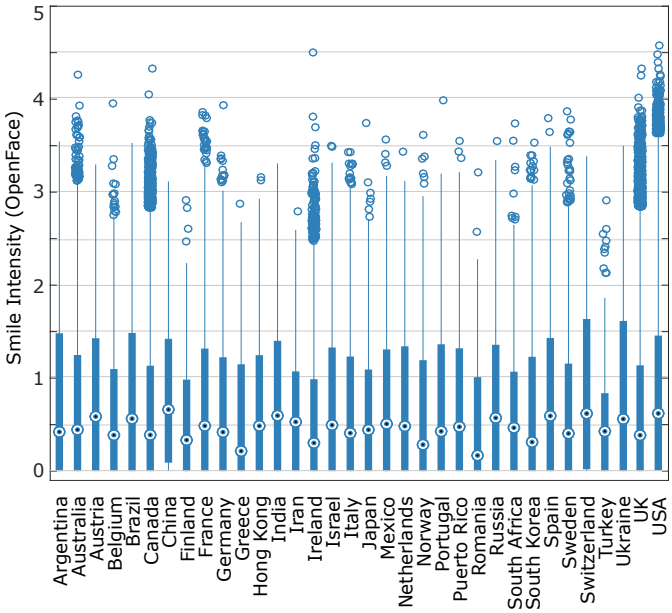

Fig. 2. Distribution of smile intensity estimates across all the images by country. Although most images contained quite subtle smiles or non-smiles, outliers with more extreme expressions were observed in most countries.

We then collected smile measurements from two trusted sources so that we could compare them to the OpenFace estimates. First, we recruited expert FACS coders to annotate the smile (AU12) intensity level in each image in the validation sample. The official FACS intensity scale has six ordinal intensity levels (0 to 5): *none*, *trace*, *slight*, *marked*, *extreme*, and *maximum*. Second, we recruited five crowdworkers to rate the positivity of the expression in each image (from 0 to 5, *neutral* to *very positive*) and the degree to which the person in each image was smiling (from 0 to 5, *not at all* to *very strongly*). We then compared the OpenFace smile intensity estimates to these measurements using correlational analyses.

Spearman correlations were used to quantify the association between continuous variables (i.e., OpenFace estimates and the average of the crowdworkers' ratings), and polyserial correlations were used to quantify associations involving ordinal variables (i.e., FACS codes) [25]. For variables measured on the same metric (i.e., ratings from different crowdworkers or measures of AU12 intensity), intraclass correlations (ICCs) were also calculated: the consistency ICC, which allows each variable to have its own mean, and the agreement ICC, which requires both variables to have the same mean [26].

IV. Statistical Modeling

Analyzing the study data presented several challenges. First, we had to account for the fact that our response variable (i.e., estimated smile intensity) was not normally distributed; rather, it was bounded between 0 and 5 and was right-skewed with many zeros. Second, we had to account for the fact that the data had a hierarchical structure with multiple images per individual and multiple individuals per country. To ignore

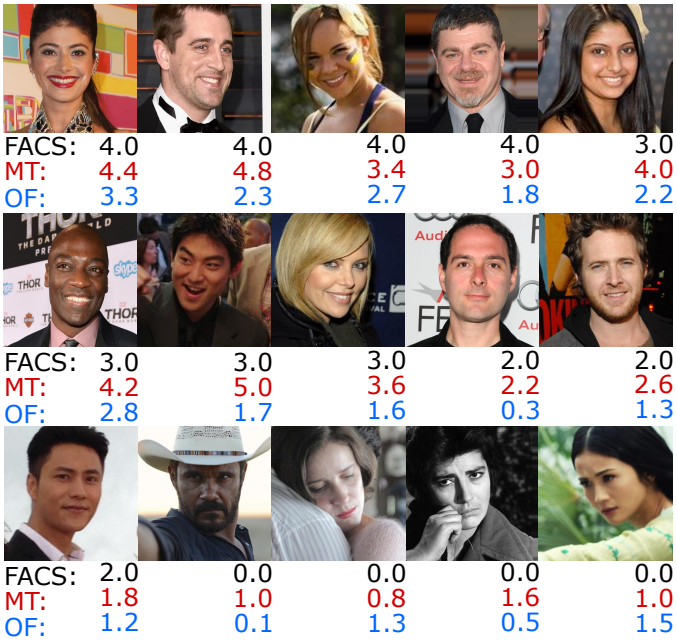

Fig. 3. Example validation images with expert AU12 coding (FACS), mean crowdworker smile ratings (MT), and OpenFace AU12 estimates (OF).

either of these challenges would likely result in a poorly fitting model and could lead to incorrect conclusions [27], [28].

To account for the non-normal distribution of the response variable, we used an extension of beta regression [29]. Beta regression is often used to model data that are measured in a continuous but bounded interval such as proportions and probabilities. The beta distribution, which beta regression uses to model the response variable, is highly flexible in that its density can take on very different shapes (e.g., unimodal, U, J, inverted J, or uniform) depending on the values of its parameters. The density of the beta distribution for $y \in (0, 1)$ is given by the following function, where B is the beta function and μ and ϕ are positive parameters that determine its shape:

$$f(y; \mu, \phi) = \frac{y^{\mu\phi-1}(1-\mu)^{(1-\mu)\phi-1}}{B(\mu\phi, (1-\mu)\phi)} \quad (1)$$

Beta regression assumes the response variable is between (but does not include) 0 and 1, so we rescaled the smile intensity estimates by dividing them by the maximum possible intensity. There were no values at 1 after rescaling, but there were many values at 0 (i.e., non-smiles). To accommodate these values, we used an extension of beta regression called zero-inflated beta (ZIB) regression [30]. This approach models the data using a mixture of two distributions: a beta distribution for the values greater than 0 and a degenerate distribution for the values at 0. The density of a ZIB distribution is given by the following function, where z denotes the probability of y being 0 and $f(y; \mu, \phi)$ refers to Equation 1:

$$f_{zi}(y; z, \mu, \phi) = \begin{cases} z, & \text{if } y = 0 \\ (1-z) f(y; \mu, \phi), & \text{if } y \in (0, 1) \end{cases} \quad (2)$$

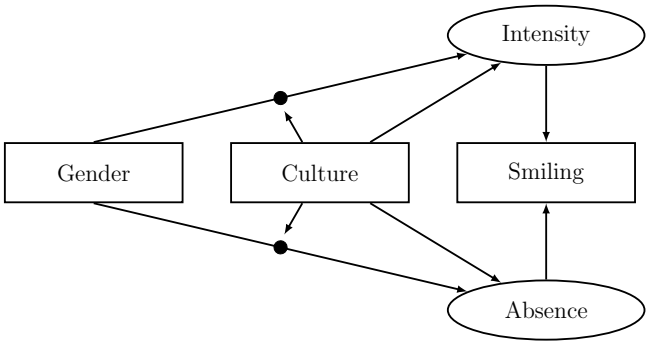

Fig. 4. Path diagram of the multilevel model. Smiling is decomposed into intensity (i.e., beta) and absence (i.e., zero-inflation) components, which are predicted by gender and culture. The random gender effects (shown as dots) are also predicted by culture, producing the gender-by-culture interactions.

One of the benefits of ZIB regression is that both the beta and the degenerate distributions can accommodate predictor variables. This allows questions to be asked about both the occurrence of the response variable (via the degenerate) and its extent or degree (via the beta). We used this ability to explain whether *and* how intensely an individual smiled.

To account for the hierarchical structure of the data, we implemented the ZIB regression model within a multilevel modeling framework [31]. Multilevel models explicitly account for the complex dependency structures of hierarchical data and allow model parameters to vary at multiple levels of the hierarchy (e.g., at the level of individuals and countries). Although there are multiple ways to implement multilevel models, Bayesian methods offer many practical and interpretive advantages over alternatives such as the ability to incorporate prior knowledge about parameters into the model and to estimate the probability of different parameter values [27], [28]. We estimated our Bayesian multilevel model using the `brms` package [32], [33] as an interface to `Stan` [34].

Our multilevel model had two levels: each data point represented a single individual from a given country. The dependency of multiple images from the same individual was handled through aggregation (i.e., by averaging all the images' smile intensity estimates). Both the non-zero smile intensity (y) and the zero-inflation probability (z) were regressed on three predictors: a dummy code representing the gender of each individual, a continuous variable representing the culture (i.e., historical heterogeneity) of that individual's country, and the cross-level interaction of gender and culture (Figure 4). The interaction allows the effects of gender and culture to be conditional on one another (i.e., for the gender effect to vary as a function of culture and vice versa).

Our goal was to gain knowledge about smiling, gender, and culture that generalizes beyond the sample of countries that were included in the analyzed data. To enable such generalizability, intercepts and slopes (e.g., gender effects) were allowed to vary across countries and the variance in these parameters (which are called random effects) was modeled. It is also possible to estimate a single value across all countries

TABLE II
BIVARIATE CORRELATIONS IN THE VALIDATION SAMPLE

Measure	OpenFace	Positivity	Smile
OpenFace AU12 Intensity			
Average Expression Positivity Rating	0.79*		
Average Smile Intensity Rating	0.78*	0.94*	
FACS Expert AU12 Intensity	0.87†	0.97†	0.94†

* Spearman correlation † Polyserial correlation

for each parameter (i.e., to omit the random effects), but doing so would limit the findings' generalizability [31].

We tried to set priors that were informative enough to rule out unreasonable parameter values without being so strong as to rule out reasonable values. This took the form of `normal(0,1)` priors for the beta regression weights, `normal(0,5)` for the degenerate regression weights, `logistic(0,1)` priors for z intercepts, `gamma(.01,.01)` priors for the ϕ parameter, half-`student_t(3,0,10)` priors for standard deviations, and `lkj(1)` priors for correlations between random effects.

The multilevel model was estimated using the No-U-Turn Sampler [35]; this algorithm converges much quicker than alternatives, especially for high-dimensional models such as multilevel models with random effects. Sixteen Markov chains were used, each with 2000 iterations, 1000 warmup iterations, and a thinning rate of 1; this setup yielded 16 000 total posterior samples. To interpret the results, we summarized these posterior distributions using medians as point estimates and highest density intervals (HDIs) as interval estimates [36]. These intervals represent the most credible values for the parameters given the data. Note that, in interpreting the results, we deliberately avoid dichotomizing them into "significant" and "non-significant" classes and instead focus on using the Bayesian posterior probability distributions to estimate each effect's magnitude and uncertainty [27], [36], [37].

V. RESULTS

A. Smile Measurement Validation

Each image in the validation sample was analyzed in four ways: OpenFace estimated the intensity of AU12, one of three expert FACS coders annotated the intensity of AU12, and five untrained crowdworkers rated the expression's positivity and smile intensity. First, we examined the inter-rater reliability of the crowdworkers using agreement ICCs. Inter-rater reliability was very high for both expression positivity, ICC(A,5) = .90, 95 % CI: [.88, .92], and smile intensity, ICC(A,5) = .90 [.88, .92]. These results suggest that the raters were using the rating scales reliably and support our use of the average of the five ratings in further analyses.

Next, we examined the correlations between the OpenFace estimates, the expert FACS codes, and the average crowdworker ratings. As shown in Table II, all of these measures were highly inter-correlated ($r > .75$), which means that the measures were very linearly related to one another and tended to increase and decrease for the same images. The measures from human sources (i.e., crowdworkers and expert FACS coders) were particularly inter-correlated ($r > .90$).

Finally, because OpenFace AU12 intensity estimates are meant to be on the same metric as FACS codes [23], we also compared them to the expert FACS codes using ICCs. Agreement was acceptable, ICC(A,1) = .71 $[.41, .84]$, but was lower than consistency, ICC(C,1) = .78 $[.73, .82]$, which suggests that the two measures had different mean levels. Visual inspection of the data suggests that OpenFace tended to underestimate the intensity of more extreme smiles.

TABLE III
STATISTICAL RESULTS WITH HIGHEST DENSITY INTERVALS (HDIS)

Regression Coefficient	Median	95 % HDI
Gender → Smile Intensity	−0.248	[−0.294, −0.205]
Culture → Smile Intensity	0.051	[0.014, 0.087]
Interaction → Smile Intensity	−0.023	[−0.046, 0.002]
Gender → Smile Absence	0.730	[0.591, 0.869]
Culture → Smile Absence	−0.145	[−0.247, −0.038]
Interaction → Smile Absence	0.077	[−0.004, 0.164]

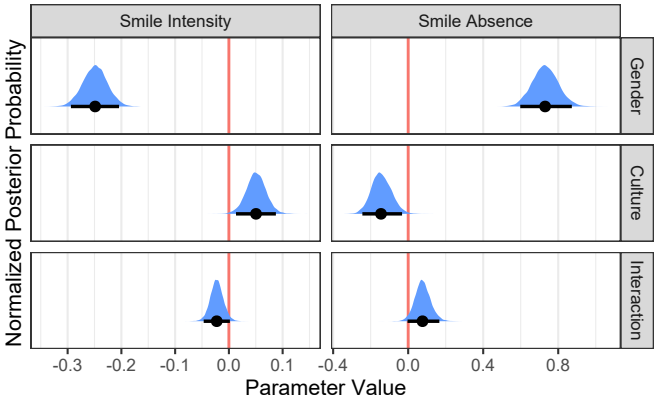

Fig. 5. Density plots for the regression coefficients where vertical height represents posterior probability and horizontal spread represents uncertainty. Dots show medians, horizontal (black) bars show 95 % highest density intervals, and vertical (red) lines highlight the value of zero (i.e., no relationship).

B. Statistical Results

The results of the multilevel model are presented in Table III and Figure 5. Results from the smile intensity component show that (1) men's smiles tended to be less intense than women's smiles $[-0.29, -0.21]$, (2) individuals from countries with higher historical heterogeneity tended to smile more intensely $[0.01, 0.09]$, and (3) this historical heterogeneity effect tended to be weaker for men than for women $[-0.05, 0.00]$. Results from the smile absence component show that (4) smiles were more likely to be absent for men than for women $[0.59, 0.87]$, (5) smiles were less likely to be absent for individuals from countries with higher historical heterogeneity $[-0.25, -0.04]$, and (6) this historical heterogeneity effect tended to be weaker for men than for women $[0.00, 0.16]$. The combination of these effects can be seen in the marginal plot (Figure 6).

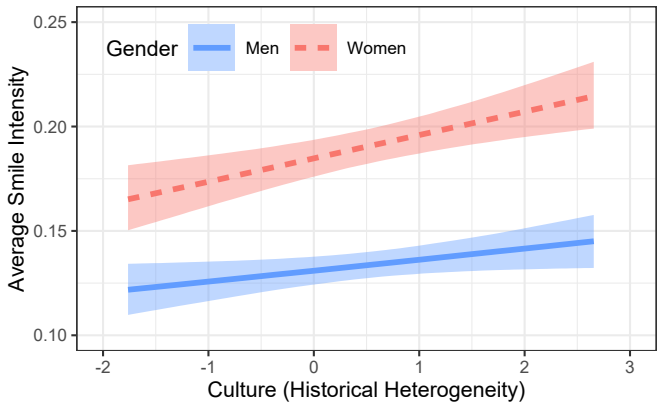

Fig. 6. Marginal effects plot with 95 % highest density intervals. Women smiled more intensely than men across countries, and historical heterogenity was positively related to smile intensity for both genders (with some evidence of a slightly larger gender difference in more historically diverse countries).

VI. Discussion

The affective computing research community has invested heavily in the development of tools for automated facial coding. These tools present a huge opportunity to advance psychological science. However, access to large video and image corpora and the necessary computational resources to analyze these are often limited to a handful of tech companies. We designed an experiment to investigate whether smaller, publicly available datasets can reveal similar insights to those obtained from large-scale analyses of privately held datasets.

Using knowledge search APIs, we collected nationality meta-data for a public dataset of celebrity face images. Using this data, we analyzed the effects of gender and culture (i.e., historical heterogeneity) on observed smile intensity. This model revealed similar effects to those observed in a previous analysis of almost one million facial videos of spontaneous behavior [1], despite being based on only 290 547 images of celebrities scraped from the Internet. We replicated evidence of both gender and cultural effects on smiling and expanded previous work by considering both the occurrence and intensity of smiling simultaneously. We also uncovered tentative evidence that the effects of gender and culture on smiling may be conditional on one another, with gender differences being slightly more pronounced in cultures with higher historical heterogeneity. These advances were enabled by use of an open-source tool for smile intensity estimation, which performed admirably in our dataset, and Bayesian multilevel ZIB regression (also implemented using open-source software).

The results of our experiment are highly promising for our original proposal. The fact that this messy and imperfect dataset was able to show evidence of the expected gender and cultural effects improves our confidence that this type of data can be representative enough to yield generalizable psychological insights, at least in some domain areas. It is also encouraging that an off-the-shelf and open-source facial coding tool was able to approximate both crowdworker ratings and expert FACS codes of smile intensity. In the spirit of democratizing insights from behavioral data, the fact that all three sources of smile measurement were highly inter-correlated bodes well for researchers with limited access to expert FACS coders. We hope that future research will continue this trend of exploring psychological questions in public behavioral data; we also hope that the researchers doing such work will invest time and effort in validating their measurements as we have tried to do in this paper. While facial coding tools can perform quite well in some settings, their accuracy cannot be taken as a given [38], [39].

There are important caveats that should be stated when analyzing data of this kind. First, the images in this dataset and other large-scale datasets are typically captured with the subject's knowledge (whether because they are posing for the press or consented to be recorded). Thus, these images do not necessarily represent fully naturalistic or spontaneous displays. Second, the contexts in which the images in these datasets were captured are very diverse. Some images are posed (e.g., professional headshots), others are action shots (e.g., a tennis player during a game), and others are acted (e.g., a still frame from a dramatized film). The fact that we observe similar cultural and gender effects as prior work suggests that display rules may be strong enough to influence behavior across contexts and that our dataset is large and diverse enough for the idiosyncratic influence of any individual image (which may be unrepresentative for many reasons) to "wash out."

As a final point of discussion, we focused in this study on images of celebrities who have chosen a life in the public eye and for whom scrutiny of their behavior is largely expected. However, there are ethical questions for the field of affective computing to wrestle with pertaining to the use of behavioral data from private individuals, such as images and videos accessible on social media. One could argue that by virtue of these data being available online, individuals have implicitly consented to their use. However, this is not a simple issue and the field will need to consider carefully how best to protect the privacy and dignity of the individuals they are studying. There is some urgency required in addressing these issues.

VII. Conclusions

Large-scale observational analyses of image and video data presents a valuable opportunity to advance psychological science. However, large image and video datasets, often collected and owned by companies, typically remain inaccessible to academics due to privacy and intellectual property concerns. We propose that interesting behavioral data already exist online (e.g., on IMDB and Wikipedia) that are publicly available and conducive to open scientific research. In the current study, we provide a proof-of-concept for this proposal by showing that convergent evidence of gender and cultural display rules about smiling can be derived from images of celebrities scraped from the internet and analyzed using facial coding software. We release this data and hope our work spurs more open exploration of this kind.

Supplemental Materials

All data, syntax, results, and figures are provided on the Open Science Framework: https://osf.io/n4grd

References

[1] J. M. Girard and D. McDuff, "Historical heterogeneity predicts smiling: Evidence from large-scale observational analyses," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 719–726.

[2] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of Nonverbal Behavior*, vol. 6, no. 4, pp. 238–252, 1982.

[3] M. Rychlowska, R. E. Jack, O. G. B. Garrod, P. G. Schyns, J. D. Martin, and P. M. Niedenthal, "Functional smiles: Tools for love, sympathy, and war," *Psychological Science*, vol. 28, no. 9, pp. 1259–1270, 2017.

[4] D. McDuff, "Smiling from adolescence to old age: A large observational study," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 98–104.

[5] W. V. Friesen, "Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules," Doctoral dissertation, University of California San Francisco, 1973.

[6] A. J. Fridlund, *Human facial expression: An evolutionary view*. Academic Press, 2014.

[7] M. LaFrance, M. A. Hecht, and E. L. Paluck, "The contingent smile: A meta-analysis of sex differences in smiling," *Psychological Bulletin*, vol. 129, no. 2, pp. 305–334, 2003.

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

[9] M. Rychlowska, Y. Miyamoto, D. Matsumoto, U. Hess, E. Gilboa-Schechtman, S. Kamble, H. Muluk, T. Masuda, and P. M. Niedenthal, "Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles," *Proceedings of the National Academy of Sciences*, vol. 112, no. 19, pp. E2429–E2436, 2015.

[10] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[11] D. Matsumoto, "Cultural similarities and differences in display rules," *Motivation and Emotion*, vol. 14, no. 3, pp. 195–214, 1990.

[12] ——, "Culture, context, and behavior," *Journal of personality*, vol. 75, no. 6, pp. 1285–1320, 2007.

[13] D. Matsumoto, S. H. Yoo, J. Fontaine, A. M. Anguas-Wong, M. Arriola, B. Ataca, ..., and E. Grossi, "Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism," *Journal of Cross-Cultural Psychology*, vol. 39, no. 1, pp. 55–74, 2008.

[14] A. H. Fischer and A. S. R. Manstead, "The relation between gender and emotions in different cultures," in *Gender and emotion: Social psychological perspectives*, A. H. Fischer, Ed. New York, NY: Cambridge University Press, 2000, pp. 71–94.

[15] A. Wood, M. Rychlowska, and P. M. Niedenthal, "Heterogeneity of long-history migration predicts emotion recognition accuracy," *Emotion*, vol. 16, no. 4, pp. 413–420, 2016.

[16] D. Matsumoto, B. Willingham, and A. Olide, "Sequential dynamics and culturally-moderated facial expressions of emotion," *Psychological Science*, vol. 20, no. 10, pp. 1269–1274, 2009.

[17] P. H. Waxer, "Video ethology: Television as a data base for cross-cultural studies in nonverbal displays," *Journal of Nonverbal Behavior*, vol. 9, no. 2, pp. 111–120, 1985.

[18] D. McDuff, J. M. Girard, and R. El Kaliouby, "Large-scale observational evidence of cross-cultural differences in facial behavior," *Journal of Nonverbal Behavior*, vol. 41, no. 1, pp. 1–19, 2017.

[19] D. McDuff, private communication, 2019.

[20] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 10–15.

[21] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[22] L. Putterman and D. N. Weil, "Post-1500 population flows and the long run determinants of economic growth and inequality," *The Quarterly Journal of Economics*, vol. 125, no. 4, pp. 1627–1682, 2010.

[23] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[24] P. Ekman, W. V. Friesen, and J. Hager, *Facial action coding system: A technique for the measurement of facial movement*. Salt Lake City, UT: Research Nexus, 2002.

[25] H. C. Kraemer, "Correlation coefficients in medical research: From product moment correlation to the odds ratio," *Statistical Methods in Medical Research*, vol. 15, no. 6, pp. 525–545, Dec. 2006.

[26] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[27] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, 3rd ed. Boca Raton, FL: CRC Press, 2014, vol. 1542.

[28] R. McElreath, *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. New York, NY: CRC Press, 2016.

[29] S. Ferrari and F. Cribari-Neto, "Beta regression for modelling rates and proportions," *Journal of Applied Statistics*, vol. 31, no. 7, pp. 799–815, 2004.

[30] R. Ospina and S. L. P. Ferrari, "Inflated beta distributions," *Statistical Papers*, vol. 51, no. 1, pp. 111–126, 2010.

[31] I. G. G. Kreft and J. de Leeuw, *Introducing multilevel modeling*. Thousand Oaks, CA: Sage Publications, 1998.

[32] P.-C. Bürkner, "Brms: An R Package for Bayesian Multilevel Models Using Stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[33] ——, "Advanced Bayesian multilevel modeling with the R package brms," *The R Journal*, vol. 10, no. 1, pp. 395–411, 2018.

[34] A. Gelman, D. Lee, and J. Guo, "Stan: A probabilistic programming language for Bayesian inference and optimization," *Journal of Educational and Behavioral Statistics*, vol. 40, no. 5, pp. 530–543, 2015.

[35] M. D. Hoffman and A. Gelman, "The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo," *Journal of Machine Learning Research*, vol. 15, pp. 1593–1623, 2014.

[36] J. K. Kruschke and T. M. Liddell, "The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective," *Psychonomic Bulletin & Review*, vol. 25, no. 1, pp. 178–206, 2018.

[37] A. Gelman, "The connection between varying treatment effects and the crisis of unreplicable research: A Bayesian perspective," *Journal of Management*, vol. 41, no. 2, pp. 632–643, 2015.

[38] J. M. Girard, J. F. Cohn, L. A. Jeni, M. A. Sayette, and F. De la Torre, "Spontaneous facial expression in unscripted social interactions can be measured automatically," *Behavior Research Methods*, vol. 47, no. 4, pp. 1136–1147, 2015.

[39] J. F. Cohn, I. O. Ertugrul, W.-s. Chu, J. M. Girard, and Z. Hammal, "Affective facial computing: Generalizability across domains," in *Multimodal Behavior Analysis in the Wild: Advances and Challenges*, X. Alameda-Pineda, E. Ricci, and N. Sebe, Eds. Academic Press, 2019, pp. 407–441.