Pre-print of article that will appear at FG 2017.

Predicting First Impressions with Deep Learning

Mel McCurrie[1], Fernando Beletti[1], Lucas Parzianello[1], Allen Westendorp[1], Samuel Anthony[2,3], and Walter J. Scheirer[1]

[1]Department of Computer Science and Engineering, University of Notre Dame
[2]Department of Psychology, Harvard University [3]Perceptive Automata, Inc.

***Abstract*— Describable visual facial attributes are now commonplace in human biometrics and affective computing, with existing algorithms even reaching a sufficient point of maturity for placement into commercial products. These algorithms model objective facets of facial appearance, such as hair and eye color, expression, and aspects of the geometry of the face. A natural extension, which has not been studied to any great extent thus far, is the ability to model subjective attributes that are assigned to a face based purely on visual judgements. For instance, with just a glance, our first impression of a face may lead us to believe that a person is smart, worthy of our trust, and perhaps even our admiration — regardless of the underlying truth behind such attributes. Psychologists believe that these judgements are based on a variety of factors such as emotional states, personality traits, and other physiognomic cues. But work in this direction leads to an interesting question: how do we create models for problems where there is no ground truth, only measurable behavior? In this paper, we introduce a convolutional neural network-based regression framework that allows us to train predictive models of crowd behavior for social attribute assignment. Over images from the AFLW face database, these models demonstrate strong correlations with human crowd ratings.**

I. Introduction

In human attribute modeling there often exists a disparity between the way humans describe humans and the way computational models describe humans. A large amount of describable attribute research in computer vision concentrates on objective traits. For example, work using the popular CelebA dataset [22], [29], [42], [40] applies different methods to model traits such as "Male" and "Bearded" with binary annotations. Beyond objective attributes, it is possible to model more subjective traits such as expression [12], [7], attractiveness [17], and humorousness [21], but research often overlooks the important interrelation between attribute modeling and social psychology. Enabling computers to make accurate predictions about objective content and enabling computers to make human-like judgements about subjective content are both necessary steps in the development of machine intelligence. Here we focus on the latter.

Specifically, we concentrate on descriptions of the face, as an abundance of social psychology research demonstrates a human tendency to make judgements in social interactions based on the faces of fellow humans [31], [38], [1]. Popular human characteristics of academic interest closely related to these social interactions include emotion [24], attractiveness [1], trustworthiness [35], [38], [28], [8], dominance [31], [24], sociability, intelligence, and morality [1]. Psychologists often specifically concentrate on trustworthiness, dominance, and intelligence because they represent comprehensive abstract qualities that humans regard in each other. Alexander Todorov, one of the foremost psychologists studying these social judgements uses dominance and trustworthiness as the basis of many in-depth studies of human judgement [35], [36], [34]. Ultimately he finds that most other recognizable subjective traits in humans can be represented as an

978-1-5090-4023-0/17/$31.00

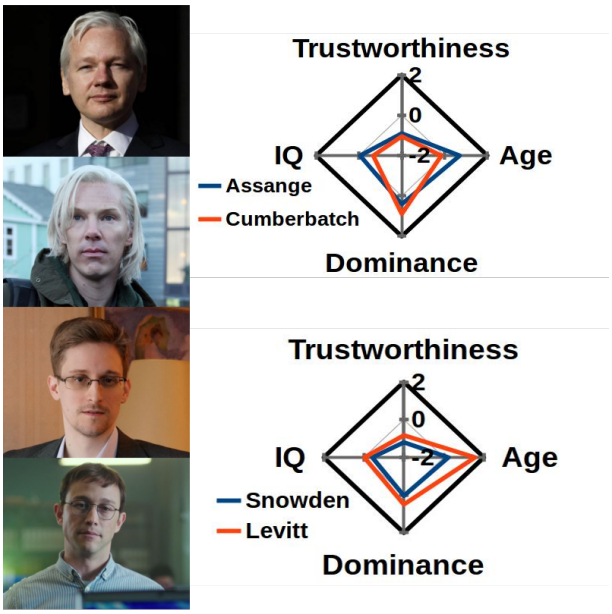


Fig. 1: Computational modeling of social attributes allows us to predict what the crowd might say about a face image. In this image we graphically compare the attribute predictions for Julian Assange and Benedict Cumberbatch, who plays Assange in the movie *The Fifth Estate*, as well as the predictions for Edward Snowden and Joseph Gordon-Levitt, who plays Snowden in the movie *Snowden*. Specifically looking at these images, our models output similar predictions between the subjects and their actors. The radar plots above reflect the output of a face processing pipeline, where faces are detected, aligned, and then processed through a deep convolutional neural network regressor that models a particular social attribute. This regression framework is the main contribution of our work. For this image we display the predictions' z-scores with respect to the training data.

orthogonal function of dominance and trustworthiness [27], which suggests these two conceptual traits are ideal for computational modeling.

Closely related to our work is research concentrated on the assessment of abstract traits in human faces based on the effect of facial contortions and positions. Inspired by animals' displays of dominance and submissiveness in respective head raises and bows, Mignault et al. specifically analyzed the effects of head tilt on the change in perceived dominance and emotion [24]. Not only does the study confirm the hypothesized disparity in perceived traits based on head tilt, but it also finds gender has a noteworthy influence on subjects' perceptions. Keating et al. assessed the effect of eyebrow and mouth gestures on perceived dominance and happiness in a cultural context [14]. The study found smiling to be a universal indicator of happiness and showed weak associations between not smiling and dominance. It also determined the effect of a lowered-brow on perceived dominance to be generally restricted to Western subjects.

In this paper we connect traditional machine learning and social psychology findings like those described above. We work specifically with traits that do not have a ground truth and can be considered abstract representations of high-level human attributes. Additionally, we introduce a convolutional neural network-based (CNN) regression framework that allows us to train predictive models of crowd behavior for social attribute assignment. Very different from prior work, we make use of a unique visual psychophysics crowdsourcing platform, TestMyBrain.org, to gather the annotations necessary for training. As a case study, we examine three purely (when analyzed in a visual context) social attributes: dominance, trustworthiness, and IQ. We also look at the more familiar objective attribute of age, but purely in the context of crowd judgements. Our models demonstrate strong correlations with crowd ratings, which we suspect are largely driven by low-level image queues.

In short, our contributions in this paper are:

- A novel ground truth-free dataset of over 6,000 images annotated for all four traits of interest.
- The deployment of a crowd-sourced data collection regime, which collects large amounts of data on high-level social attributes from the popular psychophysics testing platform TestMyBrain.org.
- The comparison of different deep learning architectures for abstract social attribute modeling.
- A set of highly effective automatic predictors of social attributes that have not been modeled before in computer vision.

II. Related Work

The related work in computer vision falls into two categories: general facial attributes, and specific CNN-based approaches. We review both in this section.

A. Attributes in Computer Vision

Due to the proliferation of low-cost high performance computing resources (*e.g.*, GPUs) and web-scale image data, large-scale image classification and labeling is now commonplace in computer vision. With respect to face images from the web, Labeled Faces in the Wild [13], YouTube Faces [39], MegaFace [26], Janus Benchmark A [15], and CelebA [22] are all popular choices for a variety of facial modeling tasks beyond conventional face recognition. Attribute prediction, where the objective is to assign semantically meaningful labels to faces in order to build a human interpretable description of facial appearance, is the particular task we concentrate on in this paper.

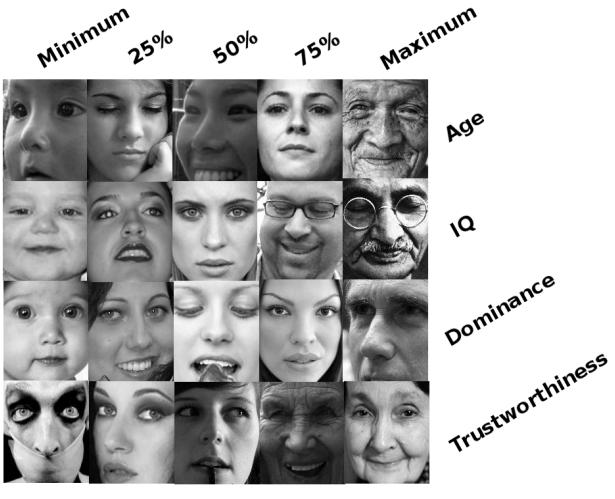
Both Farhadi et al. [9] and Lampert et al. [19] originally conceived of visual attributes as a development supporting object recognition, rather than a primary goal in and of itself. Faces, however, are a special case where standalone analysis supports applications in biometrics and affective computing. Kumar et al. used facial attributes for face verification and image search [17]. Scheirer et al. applied the statistical extreme value theory to facial attribute search spaces to create accurate multi-dimensional representations of attribute searches [30]. Siddiquie et al. modeled the relationships between different attributes to create more accurate multi-attribute searches [32]. And Luo et al. captured the inter-dependencies of local face regions to increase classification accuracy [23].

Certain traits such as Age [25], [18], [20] and gender [21], [20] have enjoyed disproportionate attention, but researchers also model numerous other facial attributes. The release of the large CelebA dataset [22] also prompted several novel studies of facial attributes on all 40 traits in the dataset [29], [42], [40]. For a comprehensive review of facial attribute work in practical biometric systems, see the review authored by Dantcheva et al. [6].

B. Convolutional Neural Networks for Attributes

Current state-of-the-art facial attribute modeling relies on CNNs. Pioneering work in the field, Golomb et al. trained a CNN with an 8.1% error rate on gender prediction [11]. More recently, Zhang et al. used CNNs alongside conventional part-based models to predict attributes such as clothing style, gender, action, and hair style from images [41]. Wang et al. applied CNNs to an automatically generated egocentric dataset annotated for contextual information such as weather and location [37]. Levi et al. used a CNN for age and gender classification from faces [20]. Liu et al. used two cascaded CNNs and trained support vector machines to separate the processes of face localization and attribute prediction [22]. And Zhong et al. extended the work of Liu et al. using off-the-shelf CNNs to build facial descriptors in a different approach to attribute prediction [42].

Most similar to our research is the recent work of Lewenberg et al. [21]. They use a CNN to predict objective traits including gender, ethnicity, age, make-up, and hair color, and subjective traits including emotional state, attractiveness, and humorousness. That research introduced a new face attributes dataset of $10,000$ images annotated for these traits. To generate this dataset, Lewenberg et al. employed Amazon's Mechanical Turk raters from the US and Canada to rate a



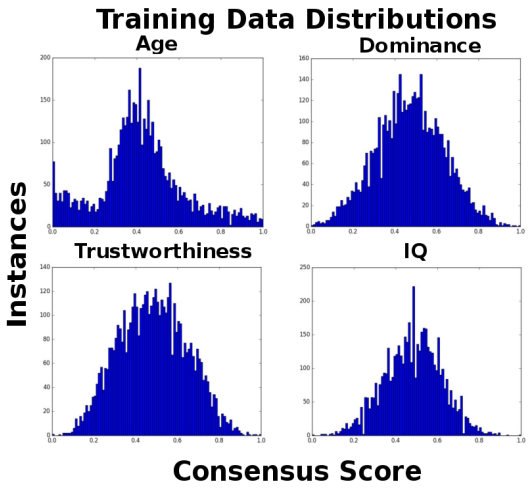


Fig. 2: We assert that to most accurately model humans' psychological judgements, each of these traits should be modeled on a continuous distribution. For this reason we employed the Likert Scale in our data collection and then took the average of human ratings for each image. This graphic shows faces at each quartile from the dataset (left) as well as the training data distributions (right), all of which seem to be close to normal.

subset of the PubFig dataset, aggregating labels from three separate individuals for each image. Notably, the work only analyzes the traits with binary classification, labeling each image as "yes" or "no" with respect to a trait. Our most immediate improvement on this work is in the way in which we collect data. We use an online psychophysics testing platform, aggregating data from a larger number of raters from an arguably more reliable and geographically variable source. In addition, we model more abstract, representational traits on continuous distributions.

Also parallel to our work, and the current state-of-the-art attribute prediction, is the work of Rudd et al. [29]. Rudd et al. employ a single custom Mixed Objective Optimization Network (MOON) to multi-task facial attribute recognition, minimizing the error of their networks over all forty traits of the CelebA dataset [22]. We use our own implementation of the MOON architecture as a basis for each separate trait in our modeling.

III. Crowd-Sourced Data Collection

In this paper we introduce a new dataset for social attribute modeling. The dataset consists of $6,300$ grayscale images of faces sampled from the AFLW dataset [16] and annotated for the four traits we study. Representative samples of the dataset for each trait can be seen in Fig. 2. This dataset is novel in that there is no ground truth. For traits such as Age and IQ, which are easy to record and described on well-known scales, it is of course possible to produce a dataset with verifiable ground truth annotations — but this is not our objective. Rather than analyze and model actual trustworthiness, dominance, IQ, and age, we choose to study people's described perceptions of the aforementioned traits. For example, our dataset does not include actual ages, instead the images are annotated by a consensus score — aggregate statistics of what many people said about the ages of the subjects in the images.

A. *TestMyBrain.org*

For this high-level, ground truth-free annotation, we use TestMyBrain.org [10], a crowd-sourced psychophysics testing website where users go to test and compare their mental abilities and preferences. It is one of the most popular "brain testing" sites on the web, with over 1.6 million participants since 2008. But what specific advantages does TestMyBrain.org have over a service like Amazon's Mechanical Turk?

TestMyBrain.org is a citizen science effort that facilitates psychological experiments and provides personalized feedback for the user, mutually benefiting both researchers and those curious about their own mind. The subject pool is geographically diverse and provides an arguably superior psychometric testing group compared to smaller more homogeneous subject pools such as that of Mechanical Turk. In addition to being an ideal setting for aggregate, cross-cultural psychometric experiments for researchers, TestMyBrain.org provides the non-monetary incentive of detailed, personalized results for subjects. Subjects visiting the site are motivated by a desire to learn about themselves and have little incentive to respond to experiments quickly or poorly. Based on these factors, we determined that the subject pool of TestMyBrain.org is ideal for the delicate task of honestly appraising abstract, ground truth-free attributes in faces.

Using TestMyBrain.org, we asked participants to judge faces for a select trait on a Likert Scale, a psychometric bipolar scaling method shown in Fig. 3. As can be seen in Table I, each face has an average of about 32 judgements for Trustworthiness and Dominance and 15 for Age and IQ. We recorded the average judgement to use as the consensus

Fig. 3: A sample behavioral task that a subject might see on TestMyBrain.org. All ratings collected for this work were on a Likert scale between 1-7, where 1 indicates the least amount of attribute presence, and the 7 indicates the most amount.

score for that image and normalized the Trustworthiness and Dominance scores. In training we map all y values so that $0 \leq y \leq 1$. We calculated the coefficient of determination (R^2) of mean human ratings from two independent sets of 943 subjects for 389 random images from the AFLW set for Trustworthiness and 400 random images from the AFLW set for dominance. The Trustworthiness R^2 is 0.93 and the Dominance R^2 is 0.88. Both of these statistics are very similar to the internal reliability calculated by Oosterhof and Todorov [27]. Thus there is indeed signal in these data that can be learned by a machine learning algorithm.

IV. CNN Regression for Social Attributes

Our algorithm is a regression model that outputs a single score from an input image. A regression, rather than a binary classification, is a more realistic representation of the initial judgements humans make. For example, from our four modeled traits, both Age and IQ are already known to be described by continuous distributions and are therefore likely judged on continuous distributions. We assert the other two modeled traits, Trustworthiness and Dominance, are similarly best described by continuous distributions. For what is discussed below with respect to architectures, assume the output is always a single floating point number from the fully-connected layers of a CNN after the convolutional layers' feature extraction.

A. Comparing Architectures: What Works Best for Social Attribute Modeling?

We initially compared five architectures with conceptually similar structures but different depths and use of regularization. We ran each with similar parameters which we determined empirically. To test very deep architectures we used the Oxford Visual Geometry Group's VGG networks [33]. We reproduced VGGNet19's convolutional architecture, modifying the shape of the input and output matrices for our smaller grayscale images and single floating point regression output. To compare results from another deep, yet slightly more shallow architecture, we also modified and used VGGNet16 in the same manner.

The newest architecture we analyzed is our implementation of the MOON architecture [29], which is more shallow than both of the VGGNet implementations. The convolutional feature extracting portion of the architecture is similar to the VGG networks in that it consists of several segments, where each segment has multiple convolutional layers followed by a max pooling layer. We modify the architecture for our smaller grayscale images and connect the convolutional layers to fully-connected layers that output a single score. With respect to our reimplementation of the MOON architecture, we made use of the Keras [5] and Theano [2] deep learning frameworks.

For comparison we also added two shallower custom architectures with fewer convolutional layers and varying regularization. In the "Shallow" network we employ three segments of convolution and max pooling connected to fully-connected layers with dropout and Parametric ReLU activations. In the "Basic6" network we employ four segments of a single convolution and max pooling followed by two fully-connected layers with ReLU activations and no dropout.

As will be discussed below in Sec. V, the differences in model performances on the validation sets during training are not very large, suggesting the architecture choice may not make a significant difference. The newer MOON architecture performs slightly better on most of the traits, however, so we chose to use it as a basis for our final optimized models. Note that the earlier work of Lewenberg et al. [21] used an AlexNet block structure augmented with supervised features (facial landmark information) and a custom loss function, while MOON is a more straightforward VGG [33] modification.

B. Hyperparameter Optimization

Rudd et al. train their MOON models on RGB images that are larger than our grayscale images and model hypothetically less abstract objective attributes from the CelebA dataset, which is annotated for binary classification. This

TABLE I: Statistics on the $6,000$ images used for training for all four social attribute classes (normalized to a $[0, 1]$ range). The "Mean Std. of Ratings" refers to the average standard deviation of the human scores for each individual image.

	Trust.	**Dom.**	**Age**	**IQ**
Mean of Ratings	0.48	0.47	0.42	0.48
Std. of Ratings	0.16	0.16	.20	0.14
Mean Std. of Ratings	0.34	0.32	0.13	0.27
Mean Num. of Ratings	32.47	32.19	15.80	15.79

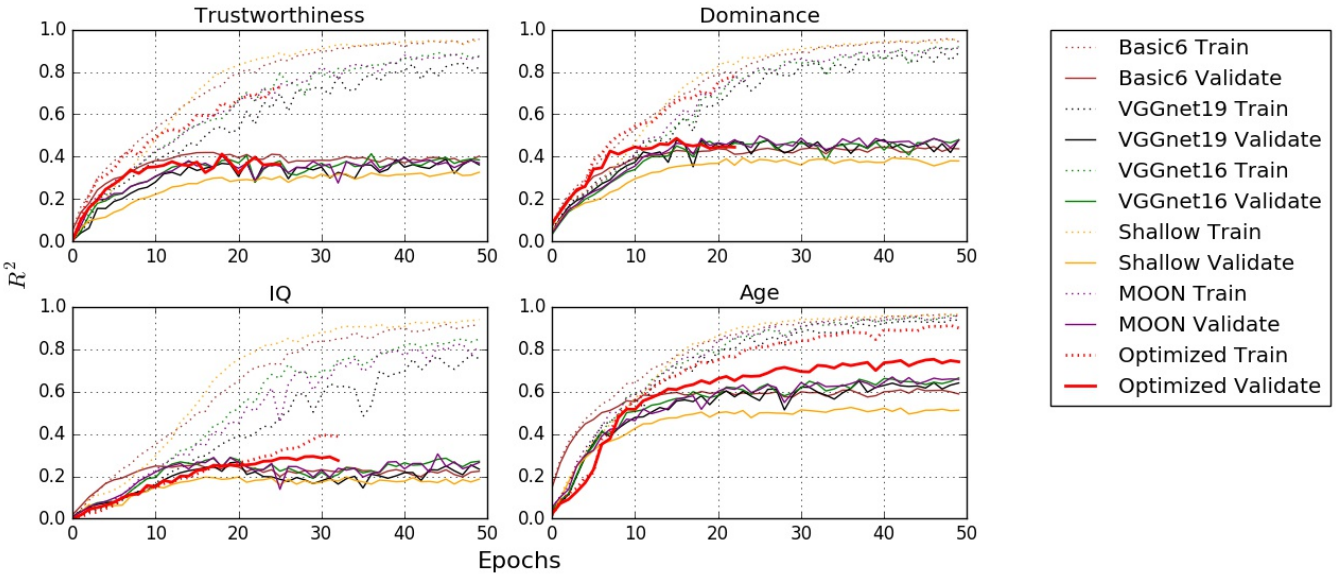


Fig. 4: In this image we compare the ability of each architecture to learn the dataset and generalize to validation data. We include all four traits, training and validation scores, and all four original architectures (best viewed in color) plus our final architecture based on the hyperparameter optimization results. Of the four original architectures the MOON models generally perform better, but our optimized models consistently perform the best. (Optimized models were trained with early stopping, as can be seen in the plots.)

suggests that our very different dataset and features could benefit from some deviations in parameter choices.

To determine the best network size and deviation in parameters from the original MOON architecture, we optimize the network for each trait using hyperopt [3], a python library for hyperparameter optimization. Our search space includes learning rate, dropout, the number of filters in each layer, the number of layers, the amount of data augmentation, and the parameters of a sampling function. Employing hyperopt with the Tree of Parzen Estimators (TPE) algorithm allowed us to test a multitude of different parameter and architecture combinations. After a very wide parameter search, we perform a refined search with early stopping, and use the best models.

We maximize the model's performance with respect to the coefficient of determination (R^2) from the regression of $\hat{y}$, the model's predicted scores, on y, the average human annotations. We use the coefficient of determination as the measure of performance because it represents the percentage of prediction variation explained by the regression model. As explained previously, our measure of performance cannot be described as "accuracy", as there is no ground truth.

As seen in Fig. 4, each trait trains very differently. Following this trend, each trait's coefficient of determination is optimized by slightly different hyperparameters and deviations from the MOON architecture as seen in Table II. However the improvements are only modest, suggesting that deeper architectures and data augmentation are not helpful for this task.

V. Experimental Evaluation

There are two important facets of evaluation with respect to our social attribute models: (1) model correlation with human crowd ratings of images, and (2) feature importance for social attribute models. Each of these facets is explored in this section. After data collection, our dataset consisted of $6,300$ grayscale images of faces, aligned to correct for in-plane rotation using the CSU Toolkit [4] and annotated for Trustworthiness, Dominance, Age, and IQ. We randomly separated 80% of the original dataset into a training set, and split the remaining 1260 images into a validation and test set (630 images each). The test set is held out during training, while the validation set is used to tune the hyperparameters.

A. Correlations with Crowd Judgements

We employ the R^2 value from a regression of $\hat{y}$, the model's predicted scores, on y, the original human annotations, as a measure of our model's performance. As seen in the supplemental material, this is a reasonable metric given the linear relationship of y and $\hat{y}$. To properly compare architectures and assess the training performance of our optimized model, we record the R^2 at each epoch and graph them in Fig. 4.

Looking at the graphs, the validation R^2 values are ultimately very similar between architectures. There is some variability in training speed, and randomly good weight initializations seem to help, however the depth of the architecture does not seem to explain any improvement in scores.

As expected, our optimized architectures outperform the other four architectures. We display our final results from the optimized networks in Table III, which shows R^2 values from

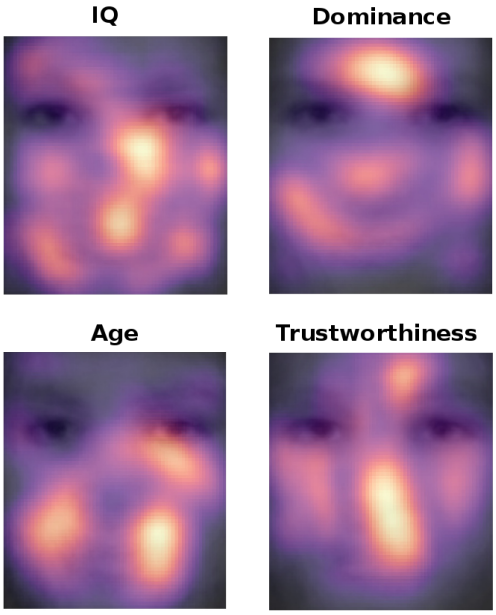


Fig. 5: We can visualize regions of the face that are most important to the trained models by systematically covering parts of the face and recording the absolute differences. Here we separately analyze 100 images of the validation set and display the average differences as a heatmap on top of the averaged faces.

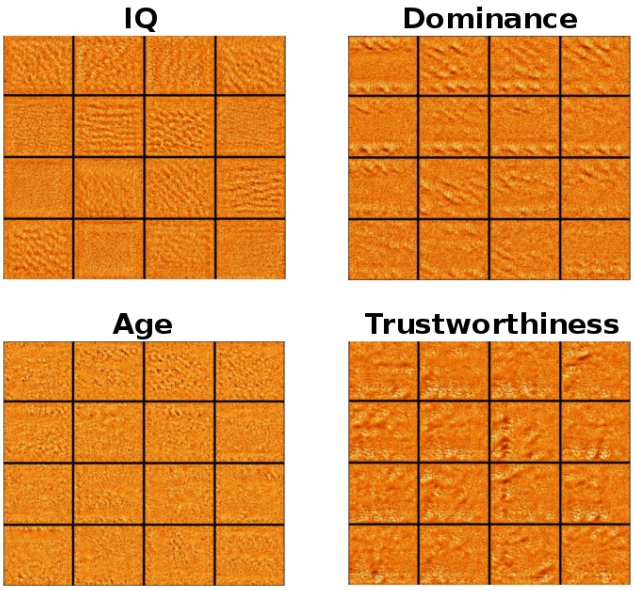


Fig. 6: Visualizing a sample of the filters from the last convolutional layer of each optimized model, we can observe the resemblance of the output to a low-level feature extractor, consistent with our observation that deeper architectures add little to no improvement. (Color added to improve contrast.)

regressions of our model's predicted values on human annotated consensus scores for both the validation and testing sets. Each trait has a slightly different coefficient of determination, however all scores are strong for a psychology-oriented experiment incorporating noisy human measurements. Our models for Age are the strongest, IQ are the weakest, and Trustworthiness and Dominance perform similarly to each other.

TABLE II: Some important hyperparameter optimization results per trait for our optimized MOON architectures.

	Trust.	**Dom.**	**Age**	**IQ**
Learning Rate	$10^{-4.2}$	$10^{-4.4}$	$10^{-4.8}$	$10^{-4.6}$
Dropout	55%	31%	45%	38%
2x Convolution 0	64	32	32	64
2x Convolution 1	64	64	128	32
2x Convolution 2	128	-	-	-
3x Convolution 3	256	256	256	256
3x Convolution 4	256	512	512	256
3x Convolution 5	256	512	512	-
FC Layers	1	3	4	3
FC Outputs	2079	2227	2187	1244

TABLE III: R^2 values of validation and testing results from our optimized MOON architectures for each trait.

	Trust.	**Dom.**	**Age**	**IQ**
Validation	.41	.49	.75	.29
Test	.38	.46	.72	.24

B. Visualizations of Feature Importance

Visualizations of the hyperparameter optimized CNN models show localized areas of importance on the face for each trait. As an example, we overlay average heatmaps for each trait on the averaged faces of 100 images from the validation data in Fig. 5. To produce these graphics we systematically moved a gray box over an image, iteratively scaling the box down after each pass. We then recorded the absolute difference in total score at each point. This visualization is intriguing because it allows us to view, in a certain image, or over an average of images, what areas of the face have the most or least significant effect on the final prediction.

Again, it is difficult to assess the validity of our models, as accuracy cannot be calculated because there is no known ground truth. Referring back to previous social psychology research [24], [14], however, both Trustworthiness and Dominance are expected to rely on the mouth. Our models indicate a heavy reliance on areas near the mouth and chin. Similarly, Keating et al. [14] determined that a lowered brow should affect the (mostly Western) perception of Dominance. Both our Dominance and Trustworthiness models approximately locate the brow mid-sections. These observations indicate that our models have learned to look in the same places that humans do, possibly replicating the way we judge high-level attributes in each other.

Another method of analyzing our models is a visualization of the filters. Our visualizations of the filters from the final convolutional layer of each network in Fig. 6 are intriguing

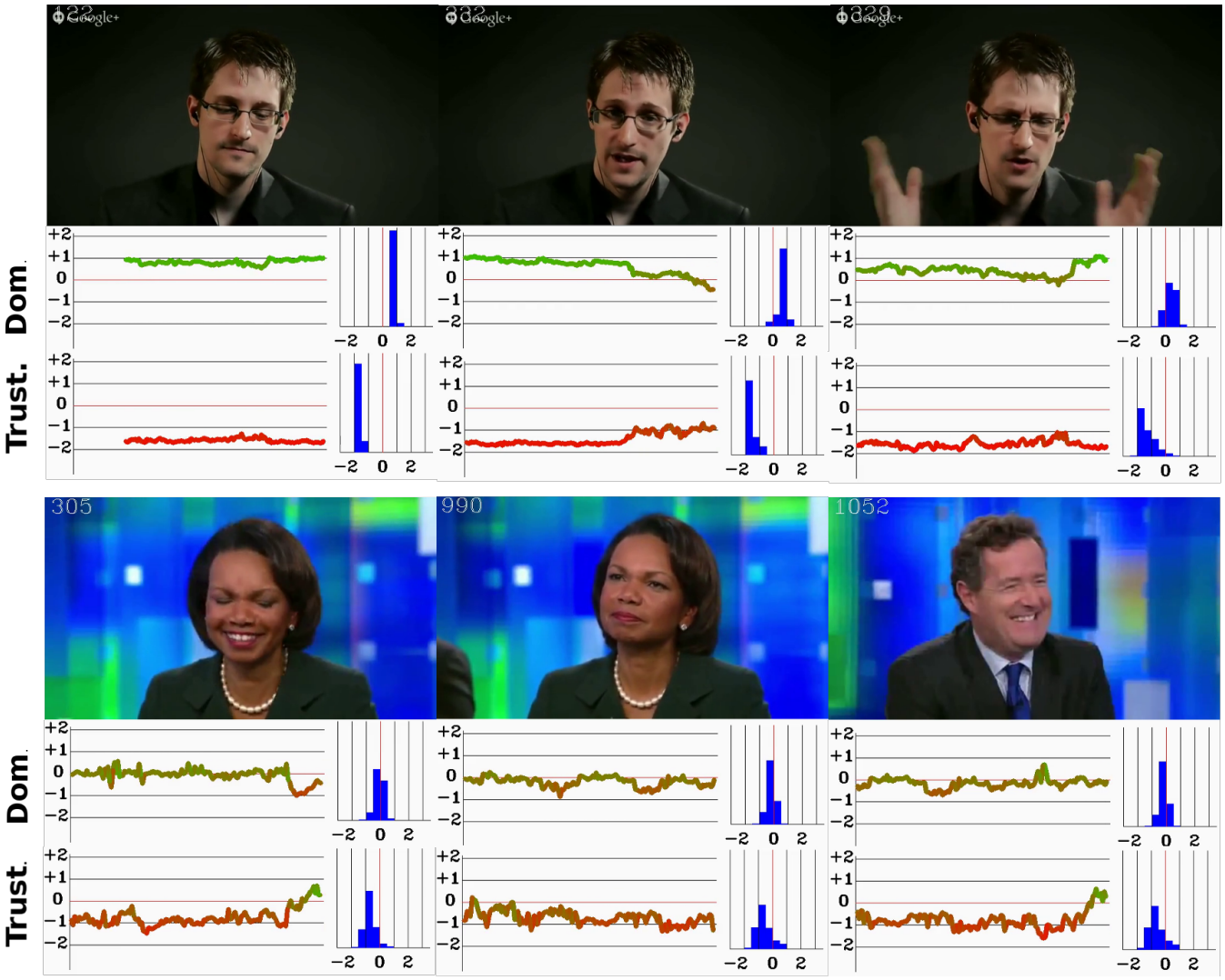


Fig. 7: Frames taken from real time video processing examples. The scores are normalized with respect to the training data statistics and then displayed over time on a line plot and a histogram. These frames exemplify changes in predictions based on facial expression and head movement.

because they resemble the output of a low-level feature extractor. This indicates that despite the high-level abstract quality of these traits, low-level features might be enough for humans to make their immediate judgements. This is consistent with our observation that deeper architectures add little to no improvement.

C. Processing Faces in Video

A very good litmus test for our models is video processing. For each frame from a video, we can apply face detection and face alignment, and then use our optimized models to predict the score of each trait. With the models loaded into memory we can even do this in real time, allowing the subjects in the videos to move and change position, simultaneously determining the change in other people's perceptions. Fig. 7 shows several frames from a couple of example videos being processed. In Fig. 7, all scores are mapped to a standard normal distribution and shown over time on both a line plot and a histogram. A selection of processed videos are provided as supplemental material.

VI. Discussion

Current state-of-the-art visual recognition algorithms in computer vision, and more specifically algorithms for facial attribute prediction [21], [29], show accuracy that promises new applications in the near future. It is in the best interest of both researchers and developers in industry to promote research that focuses on the interrelation of machine learning, computer vision, and social psychology.

A model is only as good as its data. The dataset and its annotations will ultimately have the most significant effect on the psychological validity and usefulness of the models. When annotating a dataset for subjective traits, small differences such as the number of annotators, the number of annotations, and the geographic and cultural differences of the annotators must be taken into consideration. Different cultures and languages affect the way people interpret traits,

or the description of traits. Just as intriguing as the generalizations about people that we made in our work is the study of different cultures and focus groups. Models trained only on the annotations of a focus group could generalize to new data, enabling cross-culture comparisons — useful in research, marketing, political campaigning and more.

In systematically analyzing human judgements, it is also important to choose traits that best fulfill a purpose. In our case, Trustworthiness and Dominance are the best representations of the abstract judgements humans make about each other. IQ and Age, while not as fundamental in a psychological sense, still have conceivable applications, including the assessment of preconceived notions of intelligence and seniority — subtle social cues we often take for granted.

Code, data and supplemental material for this paper can be found at: `http://github.com/mel-2445/Predicting-First-Impression`

ACKNOWLEDGEMENTS

M. McCurrie was supported by a gift from the Boeing Company. F. Beletti and L. Parzianello were supported by the Brazil Scientific Mobility Program. A. Westendorp was supported by NSF CNS RET Award #1609394. S. Anthony was supported in part by NSF SBIR Award #IIP-1621689. Hardware support was generously provided by the NVIDIA Corporation.

REFERENCES

[1] M. D. Alicke, R. H. Smith, and M. L. Klotz. Judgments of physical attractiveness: The role of faces and bodies. *Personality and Social Psychology Bulletin*, 12(4):381–389, 1986.

[2] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio. Theano: A CPU and GPU math compiler in python. In *SciPy*, pages 1–7, 2010.

[3] J. Bergstra, D. Yamins, and D. D. Cox. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *SciPy*, pages 13–20, 2013.

[4] D. S. Bolme, J. R. Beveridge, M. Teixeira, and B. A. Draper. The csu face identification evaluation system: its purpose, features, and structure. In *International Conference on Computer Vision Systems*, pages 304–313. Springer, 2003.

[5] F. Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[6] A. Dantcheva, P. Elia, and A. Ross. What else does your biometric data reveal? A survey on soft biometrics. *IEEE T-IFS*, 11(3):441–467, 2016.

[7] M. Dumas. Emotional expression recognition using support vector machines. In *International Conference on Multimodal Interfaces*, 2001.

[8] V. Falvello, M. Vinson, C. Ferrari, and A. Todorov. The robustness of learning about the trustworthiness of other people. *Social Cognition*, 33(5):368, 2015.

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *IEEE CVPR*, 2009.

[10] L. Germine, K. Nakayama, B. C. Duchaine, C. F. Chabris, G. Chatterjee, and J. B. Wilmer. Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19(5):847–857, 2012.

[11] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: A neural network identifies sex from human faces. In *NIPS*, 1990.

[12] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber, and B. Radig. Facial expression recognition with recurrent neural networks. In *International Workshop on Cognition for Technical Systems*, 2008.

[13] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller. Learning to align from scratch. In *NIPS*, 2012.

[14] C. F. Keating, A. Mazur, M. H. Segall, P. G. Cysneiros, J. E. Kilbride, P. Leahy, W. T. Divale, S. Komin, B. Thurman, and R. Wirsing. Culture and the perception of social dominance from facial expression. *Journal of Personality and Social Psychology*, 40(4):615, 1981.

[15] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *IEEE CVPR*, June 2015.

[16] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[17] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *IEEE T-PAMI*, 33(10):1962–1977, 2011.

[18] Y. H. Kwon and N. da Vitoria Lobo. Age classification from facial images. *Computer Vision and Image Understanding*, 74(1):1–21, 1999.

[19] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE CVPR*, 2009.

[20] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks. In *IEEE CVPR Workshops*, 2015.

[21] Y. Lewenberg, Y. Bachrach, S. Shankar, and A. Criminisi. Predicting personal traits from facial images using convolutional neural networks augmented with facial landmark information. *arXiv preprint arXiv:1605.09062*, 2016.

[22] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *IEEE ICCV*, 2015.

[23] P. Luo, X. Wang, and X. Tang. A deep sum-product architecture for robust facial attributes analysis. In *IEEE ICCV*, 2013.

[24] A. Mignault and A. Chaudhuri. The many faces of a neutral face: Head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*, 27(2):111–132, 2003.

[25] A. Montillo and H. Ling. Age regression from faces using random forests. In *IEEE ICIP*, 2009.

[26] A. Nech and I. Kemelmacher-Shlizerman. Megaface 2: 672,057 identities for face recognition. 2016.

[27] N. N. Oosterhof and A. Todorov. The functional basis of face evaluation. *PNAS*, 105(32):11087–11092, 2008.

[28] A. E. Pinkham, J. B. Hopfinger, K. Ruparel, and D. L. Penn. An investigation of the relationship between activation of a social cognitive neural network and social functioning. *Schizophrenia bulletin*, 34(4):688–697, 2008.

[29] E. Rudd, M. Günther, and T. Boult. Moon: A mixed objective optimization network for the recognition of facial attributes. In *ECCV*, 2016.

[30] W. J. Scheirer, N. Kumar, P. N. Belhumeur, and T. E. Boult. Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *IEEE CVPR*, 2012.

[31] C. Senior, M. Phillips, J. Barnes, and A. David. An investigation into the perception of dominance from schematic faces: A study using the world-wide web. *Behavior Research Methods, Instruments, & Computers*, 31(2):341–346, 1999.

[32] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *IEEE CVPR*, 2011.

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[34] A. Todorov, S. G. Baron, and N. N. Oosterhof. Evaluating face trustworthiness: a model based approach. *Social cognitive and affective neuroscience*, 3(2):119–127, 2008.

[35] A. Todorov and B. Duchaine. Reading trustworthiness in faces without recognizing faces. *Cognitive Neuropsychology*, 25(3):395–410, 2008.

[36] A. Todorov, M. Pakrashi, and N. N. Oosterhof. Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, 27(6):813–833, 2009.

[37] J. Wang, Y. Cheng, and R. S. Feris. Walk and learn: Facial attribute representation learning from egocentric video and contextual data. *arXiv preprint arXiv:1604.06433*, 2016.

[38] J. S. Winston, B. A. Strange, J. O'Doherty, and R. J. Dolan. Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nature Neuroscience*, 5(3):277–283, 2002.

[39] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE CVPR*, 2011.

[40] K. Zhang, L. Tan, Z. Li, and Y. Qiao. Gender and smile classification using deep convolutional neural networks. In *IEEE CVPR Workshops*, 2016.

[41] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *IEEE CVPR*, pages 1637–1644, 2014.

[42] Y. Zhong, J. Sullivan, and H. Li. Face attribute prediction using off-the-shelf CNN features. In *IAPR Int. Conf. on Biometrics*, 2016.