Title

Learning to See People Like People: Predicting Social Impressions of Faces

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 39(0)

Authors

Song, Amanda
Linjie, Li
Atalla, Chad
et al.

Publication Date

2017

Peer reviewed

Learning to See People Like People: Predicting Social Impressions of Faces

Amanda Song (mas065@ucsd.edu) †
Department of Cognitive Science
University of California, San Diego, CA 92093

Li Linjie (li2477@purdue.edu)†
Department of Computer Science, Purdue University
610 Purdue Mall, West Lafayette, IN 47907

Chad Atalla (catalla@ucsd.edu)
Department of Computer Science and Engineering
University of California, San Diego, CA 92093

Garrison Cottrell (gary@ucsd.edu)
Department of Computer Science and Engineering
University of California, San Diego, CA 92093

Abstract

Humans make complex inferences on faces, ranging from objective properties (gender, ethnicity, expression, age, identity, etc) to subjective judgments (facial attractiveness, trustworthiness, sociability, friendliness, etc). While the objective aspects of face perception have been extensively studied, relatively fewer computational models have been developed for the social impressions of faces. Bridging this gap, we develop a method to predict human impressions of faces in 40 subjective social dimensions, using deep representations from state-of-the-art neural networks. We find that model performance grows as the human consensus on a face trait increases, and that model predictions outperform human groups in correlation with human averages. This illustrates the learnability of subjective social perception of faces, especially when there is high human consensus. Our system can be used to decide which photographs from a personal collection will make the best impression. The results are significant for the field of social robotics, demonstrating that robots can learn the subjective judgments defining the underlying fabric of human interaction.
Keywords: social impression; deep learning; face perception

Introduction

With the huge success of deep learning techniques, current state-of-the-art computer vision algorithms have approached or exceeded human ability in recognizing a face (Taigman, Yang, Ranzato, & Wolf, 2014; Stewart, Andriluka, & Ng, 2016) and identifying the objective properties of a face, such as age and gender estimation, (Guo, Fu, Dyer, & Huang, 2008). However, humans not only read objective properties from a face, like expression, age, and identity, but also form subjective impressions of social aspects of a face (Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015) at first sight, such as facial attractiveness (Thornhill & Gangestad, 1999), friendliness, trustworthiness (Todorov, Baron, & Oosterhof, 2008), sociability, dominance (Mignault & Chaudhuri, 2003), and typicality. In spite of the subjective nature of social perceptions, there is often a consensus among human in how they perceive attractiveness, trustworthiness, and dominance in faces (Falvello, Vinson, Ferrari, & Todorov, 2015; Eisenthal, Dror, & Ruppin, 2006). This indicates that faces contain high-level visual cues for social inferences, therefore making it possible to model the inference process computationally. Social judgments, as an important part of people's daily interactions, have a significant impact on social outcomes, ranging from electoral success to sentencing decisions (Oosterhof & Todorov, 2008; Willis & Todorov, 2006).

Are deep learning models, which are successful in various visual tasks, also capable of predicting subjective social impressions of faces? Even before the advent of deep learning, there have been models using traditional computer vision algorithms and simulated faces to model the perception of facial attractiveness (Thornhill & Gangestad, 1999; Eisenthal et al., 2006; Kagian et al., 2008; Gray, Yu, Xu, & Gong, 2010), trustworthiness (Falvello et al., 2015; Todorov, Baron, & Oosterhof, 2008), sociability, aggressiveness (Mignault & Chaudhuri, 2003), familiarity (Peskin & Newell, 2004), and memorability (Bainbridge, Isola, & Oliva, 2013; Khosla, Bainbridge, Torralba, & Oliva, 2013). Recently, there has been work on modeling the "big five " personality traits perceived by humans when viewing another person in video clips (Escalera et al., 2016).

In this paper, we examine human social perceptions of faces in 40 dimensions extensively and systematically. We evaluate the human consistency and correlation in 40 social features (20 relevant pairs) that are typically studied by social psychologists (Todorov, Said, Engell, & Oosterhof, 2008), and relevant to social interactions (Todorov et al., 2015; Oosterhof & Todorov, 2008), and use state-of-the-art deep learning algorithms to model all 40 of them. Using the internal representations learned from the deep learning models, our model can successfully predict human social perception whenever human have a consensus. We further visualize the key features defining different social attributes to facilitate a understanding of what makes a face salient in a certain social dimension.

†These authors contributed equally.

Methods

Dataset

To predict human social impressions of faces, we use a public dataset (Bainbridge et al., 2013) consisting of 2,222 face images and annotations for 40 social attributes. Each attribute is rated on a scale of 1-9 by 15 subjects. We take the average rating from all raters as a collective estimation of human judgment for the social features of each face.

The 40 social attributes consist of 20 pairs of related traits: (attractive, unattractive), (happy, unhappy), (friendly, unfriendly), etc. Some of these traits are highly correlated and predictable from others, especially within the trait pairs. To understand the human-perceived correlations between these traits, we compute the Spearman's rank correlation between the average human ratings of every pair of social features and show their correlations in a heatmap (Figure 1(a)). We order traits in the map based on similarity and positive or negative connotation. From the figure, we see that negative social features such as untrustworthy, aggressive, cold, introverted, and irresponsible form a correlated block. Likewise, the most positive features such as attractive, sociable, caring, friendly, happy, intelligent, interesting, and confident are highly correlated with each other. Although we choose 20 pairs of opposite features, they are not completely complementary and redundant. Principal Component Analysis of the covariance matrix shows that it takes 24 principal components to cover 95% of the variance.

Regression Model for Social Attributes

After averaging human ratings, each face receives a continuous score from 1 to 9 in all social dimensions. We model these social scores with a regression model. We propose a ridge regression model on either features from deep convolutional neural networks (CNN) or traditional face geometry based features, and present results from both feature sets. Such visual features are usually high-dimensional, so we first perform Principal Component Analysis (PCA) on the extracted features of the training set to reduce dimensionality. The PCA dimensionality is chosen by cross-validation on a validation set, separately for each trait. The PCA weights are saved and further used in fine-tuning our CNN-regression model.

Regression on Geometric Features

Past studies have found that facial attractiveness can be inferred from the geometric ratios and configurations of a face (Eisenthal et al., 2006; Kagian et al., 2008). We suggest that other social attributes can also be inferred from geometric features. We compute 29 geometric features based on definitions described in (Ma, Correll, & Wittenbrink, 2015), and further extract a 'smoothness' feature and 'skin color' feature according to the procedure in (Eisenthal et al., 2006; Kagian et al., 2008). The smoothness of a face was evaluated by applying a Canny edge detector to regions from the cheek and forehead areas (Eisenthal et al., 2006). The more edges detected, the less smooth the skin is. The regions we chose to compute smoothness and skin color are highlighted in the right subplot of Figure 2. The skin color feature is extracted from the same region as smoothness, converted from RGB to HSV. However, regressing on these handcrafted features alone is not enough to capture the richness of geometric details in a face. We therefore use a computer vision library (dlib, C++) to automatically label 68 face landmarks (see Figure 2) for each face, and then compute distances and slopes between any two landmarks. Combining 29 handcrafted geometric features, smoothness, color and the distance-slope features, we obtain 4592 features in total. Since the features are highly correlated, we apply PCA to reduce dimensionality. Again, the PCA dimensionality is chosen by cross-validating on the hold out set separately for each facial attribute. Then a ridge regression model is applied to predict social attribute ratings of a face. The hyper-parameter of ridge regression is selected by leave-one-out validation within the training set.

Regression on CNN Features

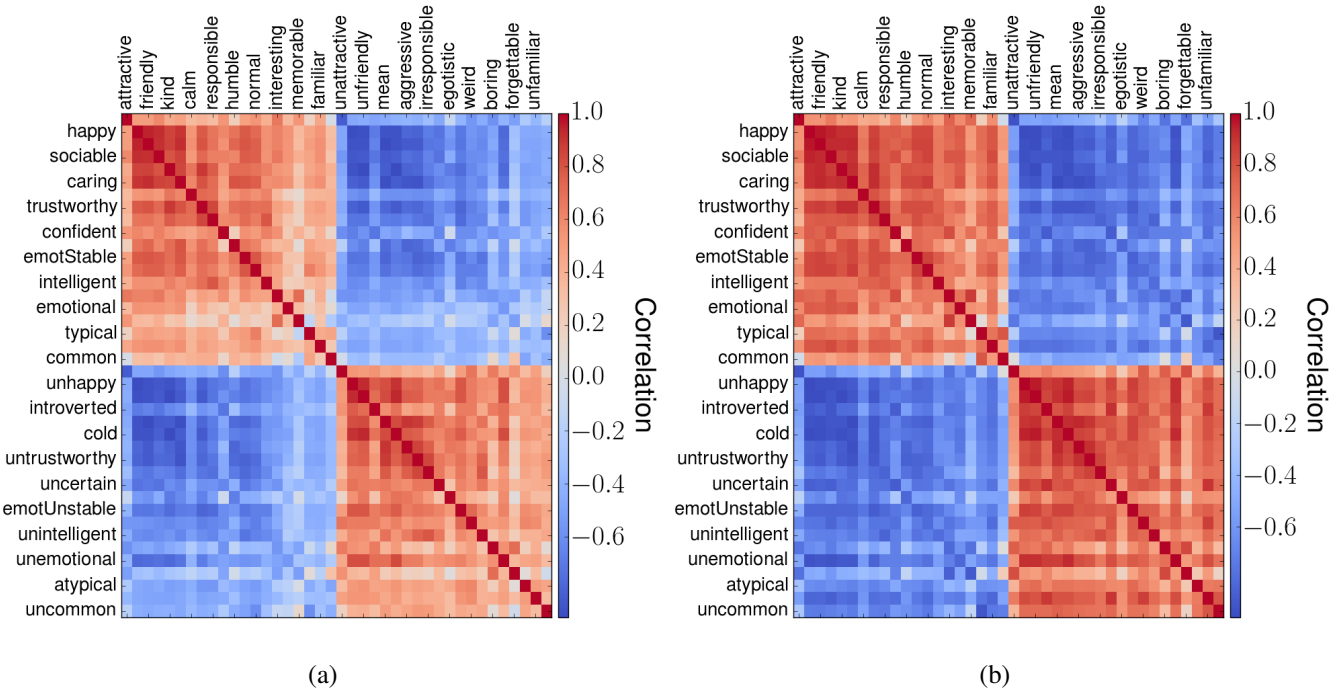
Previous studies have shown that pretrained deep learning models can provide feature representations versatile for related tasks. We therefore extract image features from pretrained neural networks, choosing from six architectures with different original training goals: (1) VGG16, trained for object recognition (Simonyan & Zisserman, 2014), (2) VGG-Face, trained for face identification (Simonyan & Zisserman, 2014), (3) AlexNet, trained for object classification (Krizhevsky, Sutskever, & Hinton, 2012), (4) Inception from Google, trained for object recognition (Szegedy et al., 2015), (5) a shallow Siamese neural network that we train from scratch to cluster faces by identity, (6) a state of the art VGG-derived network (Face-LandmarkNN) trained for the face landmark localization task.

To find the best CNN features among the six networks, we first find the best-performing feature layers of each network in the ridge regression prediction task. Before the ridge regression, we perform PCA and pick the PCA dimensionality that gives best results on the validation set. Then, we compare the results among networks to select the best features overall.

Results

After comparing all 6 networks, we find that the conv5_2 layer of VGG16 (trained for object classification) lead to the best results. This set of features significantly outperforms the three networks trained solely on faces, while also slightly outperforming AlexNet and Inception networks. These best-performing CNN features also exceed the prediction correlation of the geometric features in most attributes. Figure 3 compares prediction performance of the CNN model and the geometric feature model.

We speculate that the poor performance from the face recognition networks can be attributed to their optimization for specific facial tasks. Learning face landmark configurations and differences between faces that define identity may not correlate well with the task at hand, which looks for commonalities behind certain social features beyond identity. The



(a) (b)

Figure 1: Correlation heatmaps among social features. (a): human; (b): CNN-based model.

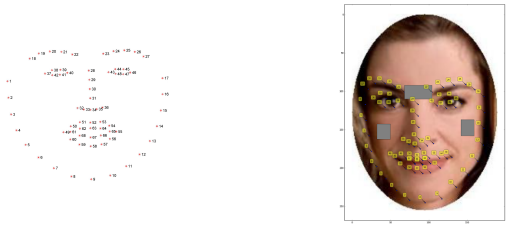


Figure 2: 68 face landmarks labeled by dlib software automatically. The gray regions are used for computing smoothness and skin color.

landmark networks should presumably give results similar to the geometric features, but did not learn features corresponding to all of the features we manually extracted.

We also try fine-tuning the best performing CNN model with back propagation but do not observe further improvement in performance. Hence our reported results are without fine-tuning.

To evaluate model performance, we did a random train/validation/test split 50 times, with a ratio of 64/16/20 respectively. The prediction performance of our model is evaluated using Pearson's correlation with the average human ratings on the test set. For each social attribute, we also compute human group consistency as an index of the strength of learning signal.

Among the social attributes, human subjects agree most about 'happy' and disagree most about 'unfamiliar.' For both regression models (CNN based regression and geometric feature based regression), model performance grows as the consensus on a social trait increases.

Since a change in expression would produce a change in landmark locations, it is not surprising that landmark-based geometric features achieve comparable or slightly higher correlation with the CNN model when predicting social attributes which are highly related to expressions (such as 'happy', 'unhappy', 'cold' and 'friendly' etc). For other social attributes, the CNN model performs better, by about 0.04 higher in correlation on average. This implies that CNN features encode much more information than landmark-based features. It is useful to visualize such features to understand what aspects make them powerful enough to predict social attributes.

Evaluating Against Human Consensus

An important gauge of model success is quantitative comparison between the subjective social features predicted by our best performing model and those perceived by humans. We take our model predictions, compute the Spearman correlation between every pair of traits, and display them in a heatmap (see Figure 1 (b)). The resulting heatmap shares similar patterns with the figure generated from average human ratings (see the left panel in Figure 1). Pearson Correlation between the upper triangle of the two similarity matrices (human and model prediction) is 0.9836. This suggests that our model successfully preserves human-perceived relationships between traits.

Since these social impressions are subjective ratings, it is informative to examine the extent with which people agree

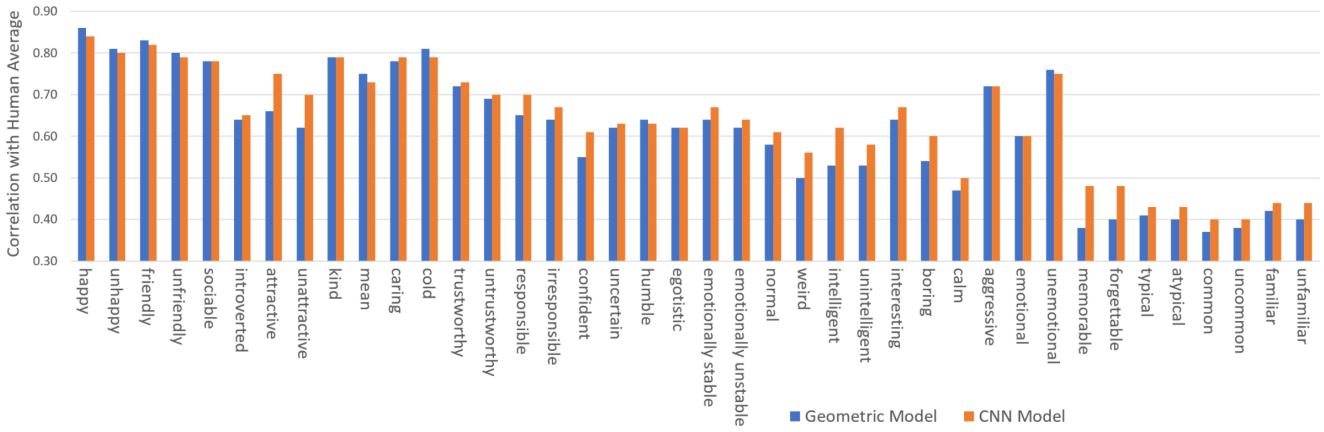


Figure 3: Model comparison on 40 social features.

with each other on these judgments. To calculate human group consistency, we perform the following procedure 50 times for each attribute and then average the results: (1) For each face, we randomly split the 15 raters into two groups of 7 and 8. (Note: The raters assigned to each face are generally different sets). (2) We calculate the two groups' average ratings for each face, obtaining two vectors of length 2,222 (the number of faces in the dataset). (3) Finally, we calculate the Pearson correlation between the two vectors. We find that human agreements covary with model performance and observe an extremely high correlation, as illustrated in Figure 4.

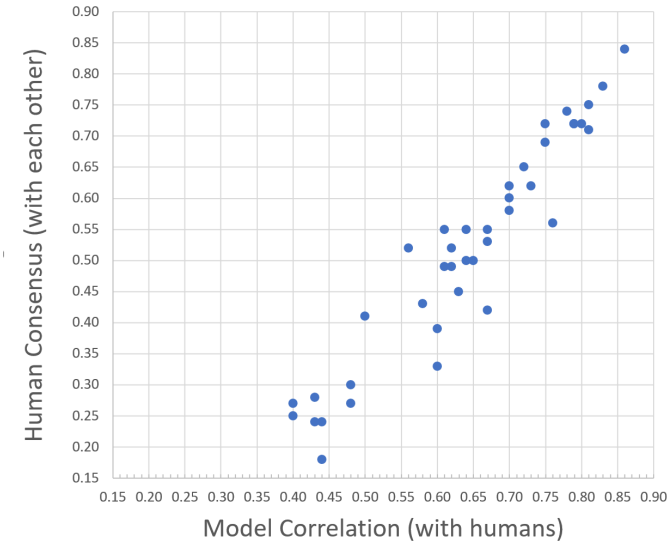


Figure 4: Human within group consistency vs. model's correlation with human average. Pearson correlation $\rho = 0.98$, $p < 10^{-5}$

Feature Visualization

Here, we visualize features from our model which are important for social perceptions. We choose facial attractiveness as an example, but the same method can be applied to the other social features.

To identify visual features that ignite attractiveness perception, we find the top 9 units of highest influence on attractiveness at conv5_2 as follows. First, we compute a product of three terms: (1) A unit's activation from conv5_2, (2) that unit's weight to the following fc_PCA layer, (3) the fc_PCA unit's weight to the output unit. We then sort all conv5_2 units' average products of these three terms and identify the top 9 neurons that contribute to the output neuron for the corresponding social feature. Then we employ the method described in (Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015; Zeiler & Fergus, 2014) to find top-9 input images that cause high activations in each of the top-9 conv5_2 neurons. Also we use deconvolution to create an image of the features activating that unit for each face, with varying levels of success.

Figure 5 captures the features that are important for predicting the attractiveness of a face. The feature importance descends from left to right and top to bottom. The important features identified by our model are related to eyes, hair with bangs, high nose-bridges, high cheeks, dark eyebrows, strong commanding jawlines, chins, and red lips. Note that among the 9 cropped input image patches, not all the faces are perceived as attractive overall; despite having a feature that contributes to attractiveness. An attractive face needs to activate more than one of these features in order to be considered attractive. This observation agrees with our intuition that attractiveness is a holistic judgment, requiring a combination of multiple features.

It also seems that several attractiveness features include relationships between different facial features. For example, while the first feature in the upper left of the figure emphasizes the eye, it also includes the nose. This is also true of the upper right feature. Additionally, smiling is important in perceived attractiveness, as emphasized by the feature in the lower left of the figure.

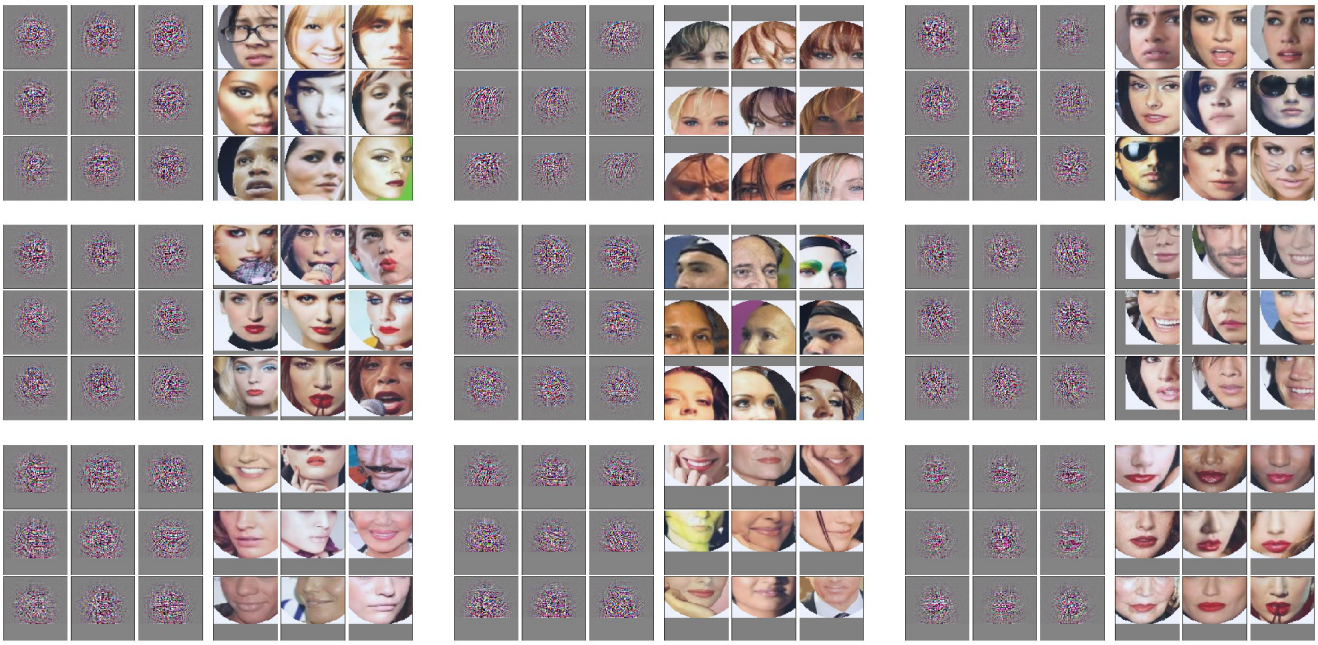


Figure 5: Visualization of features in the pretrained-VGG16 regression network. For conv5_2 layer, we show the top 9 activations of the top 9 neurons that maximally activate the attractiveness neuron across the training data, projected down to pixel space using the deconvolutional network approach (Zeiler & Fergus, 2014) and their corresponding cropped image patches. Best viewed in electronic form, and zoomed in.

Conclusion

We have shown that a deep network can be used to predict human social perception of faces, achieving high correlation with the average human ratings. As far as we know, this is the widest exploration of social judgment predictions, showing human-like perceptions on 40 social dimensions. Reflecting previous work in recognizing facial expressions, where happiness is the easiest to recognize, our highest correlation is on the happy feature. However, previous work in this area tends to classify a face as happy or not, rather than the degree of rated happiness. By predicting this as a continuous value, rather than categorical data, the subjective nature of human judgment is modeled smoothly, along with the subjective face trait landscape.

We find that, for attributes which are recognized via facial actions, such as happy, unhappy, or aggressive (probably associated with anger) or lack of facial action, such as cold or unemotional, a simple regression model based on the placement of facial landmarks works well, although the deep network performs nearly as well.

Of greater significance is our model's correlations with human judgments for traits such as trustworthiness, responsibleness, confidence, and intelligence, which correspond to more static features of the face. In this area, the deep network, which responds to facial textures and shape, has superior performance. While these judgments do not correspond to the traditional notion of "ground truth", they are descriptions for which humans have a fair amount of agreement, suggesting the presence of a signal to be recognized.

Furthermore, we have shown, yet again, that a machine can recognize attractiveness. For this dataset, our deep network correlates with average human ratings at 0.75. This provides a new benchmark for this dataset. This is one of a few areas where the deep network significantly outperforms the geometric features, as skin texture is likely to matter.

Many of these features are redundant. For example, friendly and happy are highly correlated (see Figure 1, and the red block indexed by happy and friendly). Similarly, aggressive and mean are highly correlated, which presumably requires *not* smiling. Meanwhile, it is also noteworthy that some traits considered to be "opposite" in this list are not simply the inverse of one another. For example, there is a large difference in human agreements on "sociable" (0.74) versus "introverted" (0.50), suggesting they are not opposites.

We also examined some of the features from the deep network. It is notable that these are difficult to verbalize, which is quite different from geometric features.

These results are significant for the field of social robotics. While a robot should not purely judge a human on appearance, much of human interaction is dictated by the underlying fabric of social impressions. Thus, it is important for a robot to be aware of this subjective social fabric, opening the door to useful knowledge such as whether humans might judge a person to be trustworthy. These judgments may happen subconsciously for humans, while a robot can be more objective, predicting these judgments and objectively choosing when to

consider them in a decision. A robot need not treat an attractive or unattractive person differently for its own purposes, but this knowledge could affect how interactions are made for the sake of the human, knowing in advance how that person may feel that they fit into the social landscape.

Expansions on this work may include investigating the image properties that determine high level social features, beyond the attractiveness features we display in Figure 5. Additionally, social trait prediction may benefit from a single model with a shared representation, while this paper approaches each attribute as a separate regression task.

For future work, we aim to develop a generative model which can automatically modify a face's attributes (either objective or subjective) while preserving its realism and identity. Practically speaking, such a model could improve a face's perceived social features in positive ways (e.g. make a face look more sociable, trustworthy). More importantly, it would enable psychologists to quantify human biases during the formation of social impression in a precise and systematic manner. Psychologists could generate variants of a real face differing in age, gender, race, and explore how various factors separately and jointly affect the social impressions of faces.

References

Bainbridge, W. A., Isola, P., & Oliva, A. (2013). The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, *142*(4), 1323.

Eisenthal, Y., Dror, G., & Ruppin, E. (2006). Facial attractiveness: Beauty and the machine. *Neural Computation*, *18*(1), 119–142.

Escalera, S., Torres Torres, M., Martinez, B., Baró, X., Jair Escalante, H., Guyon, I., ... others (2016). Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–8).

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

Gray, D., Yu, K., Xu, W., & Gong, Y. (2010). Predicting facial beauty without landmarks. In *The European Conference on Computer Vision (ECCV-2010)* (pp. 434–447). Springer.

Guo, G., Fu, Y., Dyer, C. R., & Huang, T. S. (2008). Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, *17*(7), 1178–1188.

Kagian, A., Dror, G., Leyvand, T., Meilijson, I., Cohen-Or, D., & Ruppin, E. (2008). A machine learning predictor of facial attractiveness revealing human-like psychophysical biases. *Vision Research*, *48*(2), 235–243.

Khosla, A., Bainbridge, W. A., Torralba, A., & Oliva, A. (2013). Modifying the memorability of face photographs. In *International Conference on Computer Vision (ICCV-2013)* (pp. 3200–3207).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, *47*(4), 1122–1135.

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

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.

Peskin, M., & Newell, F. N. (2004). Familiarity breeds attraction: Effects of exposure on the attractiveness of typical and distinctive faces. *Perception*, *33*(2), 147–158.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stewart, R., Andriluka, M., & Ng, A. Y. (2016, June). End-to-end people detection in crowded scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2016).*

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2015)* (pp. 1–9).

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2014)* (pp. 1701–1708).

Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences*, *3*(12), 452–460.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience*, *3*(2), 119–127.

Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Reviews of Psychology*, *66*(1), 519.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460.

Willis, J., & Todorov, A. (2006). First impressions making up your mind after a 100-ms exposure to a face. *Psychological science*, *17*(7), 592–598.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV-2014)* (pp. 818–833).