From prediction to modification - modeling first impression of faces

Amanda Song (mas065@ucsd.edu) †
Department of Cognitive Science

Chad Atalla (catalla@ucsd.edu) †
Department of Computer Science and Engineering

Bartholomew Tam (b4tam@ucsd.edu)
Department of Electrical and Computer Engineering

Li Linjie (linjieli222@gmail.com)
Department of Electrical and Computer Engineering

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, 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 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. This illustrates the learnability of subjective social perception of faces, especially when there is high human consensus. To verify the generalization ability, we apply the model on a large dataset, CelebA, and empirically verify the quality of model predictions. To further probe what makes a face salient in certain traits, we develop ModifAE: a novel standalone autoencoding neural network that can learn to make continuous modifications on multiple traits. We train ModifAE to modify continuous first-impression face traits, from our predicted dataset, and empirically show that this modification network produces convincing modifications, demonstrating the accuracy of the predictive model. Both the prediction and modification networks have wide applications in real life.

Keywords: social impression; autoencoding; face perception

Introduction

Humans are skilled at parsing information from faces. Apart from making objective inferences such as identity, age, and gender of a person, humans also form first impressions of a face, such as facial attractiveness, friendliness, trustworthiness, sociability, and dominance. In spite of the subjective nature of these impressions, there is often a consensus among human in how they perceive attractiveness and trustworthiness in faces. This indicates that faces contain high-level visual cues for social inferences, therefore making it possible to model the inference process. Given the profound social outcomes (electoral success, sentencing decisions) resulted from these subjective judgments(Todorov, Olivola, Dotsch, & Mende-Siedlecki, 2015), it is crucial to understand the nature of these social impressions.

In this paper, we examine human social perceptions of faces extensively and systematically. We evaluate human consistency in 40 social features that are typically studied by social psychologists (Todorov et al., 2015). Then, using internal representations from deep learning models, we propose a computational model that can successfully predict human social perception for traits where humans have consensus. This model can generalize well to an entirely new dataset. Lastly, we develop an image modification model - ModifAE, which can modify multiple social impression attributes of faces.

Dataset

We use a dataset (Bainbridge, Isola, & Oliva, 2013) consisting of 2,222 face images and annotations for 20 pairs of social attributes. Each attribute is rated by multiple subjects. We take the average rating as the group opinion. 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)). 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.

Prediction Model for Social Impressions

After averaging human ratings, each face receives a continuous score in all social dimensions. We model these 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.

† These authors contributed equally.

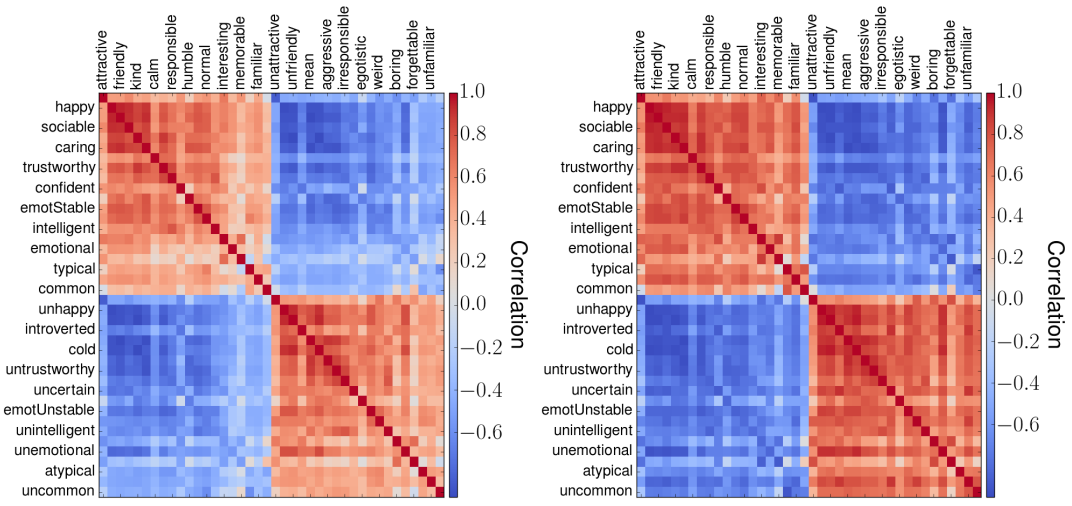

Figure 1: Correlation heatmaps among social features. (a): human; (b): CNN-based model.

Regression on Geometric Features

Past studies have found that geometric configurations of a face can predict facial attractiveness. To test if other social attributes might also be predicted 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. We also detect 68 face landmarks for each face, and then compute distances and slopes between every two landmarks. Combining 29 handcrafted geometric features, smoothness, color and the distance-slope features, we obtain 4592 features in total. PCA is applied to reduce dimensionality. 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 versatile feature representations. Therefore, we extract image features from pretrained neural networks, choosing from six architectures with different original training goals: face identification, object classification, face landmark detection.

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 of Prediction Model

After comparing 6 neural networks' all layers, we find that the conv5_2 layer of VGG16 (trained for object classification) lead to the best results. These best-performing CNN features also exceed the prediction correlation of the geometric features in most attributes. Figure 2 compares prediction performance of the CNN model and the geometric feature model.

We speculate that the poor performance from the face recognition networks is due to their optimization for specific facial tasks. Learning face landmark configurations or 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.

To evaluate model performance, we did a random train/validation/test split. The prediction performance is evaluated using Pearson's correlation on the test set. For each social attribute, we also compute human group consistency as an index of the strength of learning signal.

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 when predicting social attributes that are highly related to expressions (such as 'happy', 'unhappy', 'cold' and 'friendly'). For other social attributes, the CNN model performs better, suggesting that CNN features encode much more information than landmark-based features.

Evaluating Against Human Consensus

To gauge model success, we conduct a quantitative comparison between the impressions 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 pre-

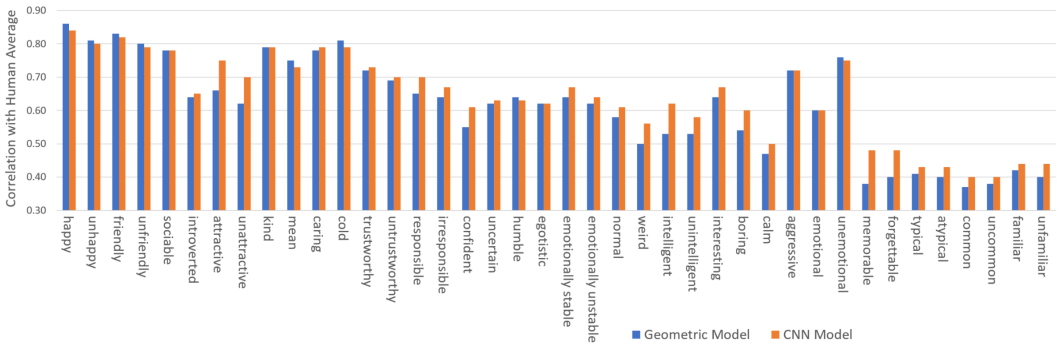

Figure 2: Model comparison on 40 social features.

diction) is 0.9836. This suggests that our model successfully preserves human-perceived relationships between traits.

To examine human consensus level, we calculate human group consistency following the procedure in (Bainbridge et al., 2013). We find that human agreements covary with model performance and observe an extremely high correlation, ρ = 0.98, $p < 10^{-5}$.

Generalize to a Larger Dataset

To test if our prediction model generalize beyond the relatively small dataset, we use it to make new predictions on CelebA dataset(Liu, Luo, Wang, & Tang, 2015), which consists of over 200,000 images of celebrities. Example faces and their predicted ratings are shown in Figure 3.

We ran an AMT experiment to check how our predicted values align with human in aggressive, responsible, trustworthy and emotional judgments. For each trait, we pick up 40 pairs of images. Among each pair, one of the image receives a high score, the other receives a low score, as predicted by our model. 30 AMT workers are asked which face better exemplifies the specified attribute. We than calculated the overall likelihood that the face of higher predicted score is chosen by workers for each attribute. As seen in Table 1, all the attributes predicted by the prediction network align well with human judgments. Thus, we have verified that our model generalizes well to the new dataset.

Figure 3: Examples of predicted impressions of CelebA faces.

Table 1: Verification of model generalization in CelebA dataset

Attribute	Chose "correct" member of the pair
Aggressive	0.9509
Emotional	0.9234
Responsible	0.7783
Trustworthy	0.8780

From Prediction to Modification

To demonstrate the quality of predicted impressions and visualize what facial dimensions are vital for social impressions, we train an image modification network with predicted social impression ratings of CelebA (Liu et al., 2015) images. We develop a special image modification network: ModifAE that can make continuous modifications of first impressions for faces.

Architecture The ModifAE architecture (Figure 4) consists of two autoencoding paths (traits and image) which fuse in the middle of the network. Two FC layers project the traits to the image feature map size at the bottleneck. Then, the values in the image feature maps are multiplied by the projected trait values. Next, we reduces the bottleneck to 16 filters, using 1×1 convolutions, fusing together the image and trait information. Then, those values are used to predict the trait output through another FC layer with linear activation.

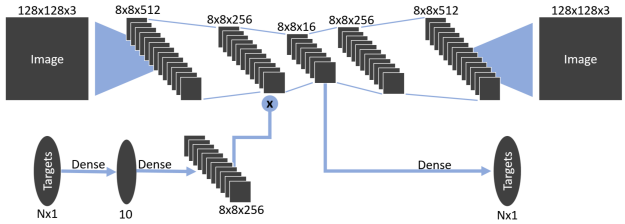

Figure 4: General illustration of ModifAE architecture.

Training ModifAE is trained soley on an autoencoding task. At training time, the objective is to take in an image and its

traits, autoencode both, and arrive at outputs identical to the inputs. No form of adversarial or cycle training is necessary. Despite this, the trained network can modify images without obscuring identity traits.

Qualitative Evaluation

Here, we show that ModifAE successfully makes continuous modifications on multiple traits (see Figure 5). We trained ModifAE on two traits: "attractive" and "emotional." The picture in the upper left corner is the original, with its true trait values next to it. Looking at the (0,0) point in results (unattractive and unemotional) her prominent cheekbone appears to sag, and her smile becomes a frown. In general, as she becomes more emotional, her smile increases, and as she is made more attractive, her smile increases and skin becomes smoother. These modifications give us an easily interpretable window into what the predictive model considers when rating faces.

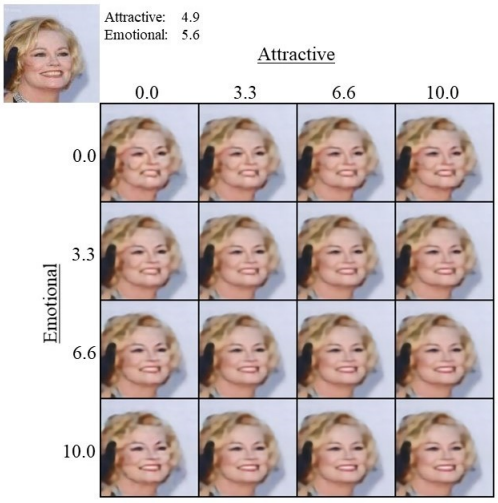

Figure 5: Image modifications by ModifAE.

Quantitative Evaluation

We assessed human interpretations of the modified faces in two traits: "trustworthy" and "aggressive." For each trait, participants recruited from AMT see a sequence of 120 image pairs. They are asked to pick which image most exemplifies the trait in every pair. Each sequence contains 10 ground truth pairs and 110 modified pairs. We calculate the fraction of subjects that chose the image with the increased trait. This is shown in Table 2, along with the human accuracy with the ground truth examples. In the "aggressive" case, our modified images achieve 79.9% of the consensus that the ground truth pairs achieve. For "trustworthiness," the modified images achieve 81.1% of the consensus on true images. It shows that the predictive model captures robust information about the examined traits and successfully passed that information to the modification model.

Table 2: Human evaluation of modified images

Attribute	Performance	
Aggressive	Ground Truth	0.9363
	Modified Faces	0.7484
Trustworthy	Ground Truth	0.8571
	Modified Faces	0.6959

Conclusion

We have shown that a deep network can be used to predict human first impressions of faces, achieving high correlation with the average human ratings. In addition, our predictive model can generalize to an entirely new dataset that mimics real life scenarios. We further employ a generative model, ModifAE, to automatically modify a face's attributes while preserving its realism. Successful modifications by ModifAE demonstrate the accuracy of the predictive model in addition to the power of the modification model. Both models have wide applications in real life as well as in academics. For instance, the prediction model can guide people to manage their impressions by selecting photos which exemplify certain desired attributes. With the modification model, psychologists can generate various realistic faces with precise control of social impressions.

Acknowledgments

This work was supported by Guangzhou Science and Technology Planning Project (201704030051).

References

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

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (iccv)*.

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.

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.