UNIVERSITY OF CALIFORNIA SAN DIEGO

A Data-Driven Study of Cross-Cultural Social Impressions on Faces

A Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Computer Science

by

Weifeng Hu

Committee in charge:

Professor Gary Cottrell, Chair
Professor Lawrence Saul
Professor Ed Vul

2020

The Thesis of Weifeng Hu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Garrison Cottrell for his support as the chair of my committee. His guidance has proved to be valuable and his funding helps make the data collection possible.

I would like to acknowledge Professor Edward Vul for his support on my thesis. His expertise knowledge has made our data analysis possible.

I would like to acknowledge Amanda Song of our research team for her support on my thesis. Her ideas have led to our excellent works.

Chapter 1 and 3, in full, is currently being prepared for submission for publication of the material. Song, Amanda; Hu, Weifeng; Yadav, Pratap Devendra; Wen, Fangfang; Zuo, Bin; Cottrell Garrison; Vul, Ed. The dissertation/thesis author was the primary investigator and author of this material.

Chapter 4, in full, has been submitted for publication of the material as it may appear in Advances in Neural Information Processing Systems 33, 2020, Hu, Weifeng; Yadav, Pratap Devendra; Song, Amanda; Vul, Ed.; Cottrell Garrison. The dissertation/thesis author was the primary investigator and author of this paper.

VITA

2018	Bachelor of Science in Computer Science, *Summa Cum Laude*, University of Michigan, Ann Arbor
2019-2020	Graduate Teaching Assistant, University of California San Diego
2020	Master of Science in Computer Science, University of California San Diego

PUBLICATIONS

Weifeng Hu*, Amanda Song*, Devendra Pratap Yadav, Fangfang Wen, Bin Zuo, Ed Vul, and Garrison Cottrell. “Do you see what I see? A cross-cultural comparison of social impressions of faces”. In Proceedings of the 42nd Annual Conference of the Cognitive Science Society, 2020. (* equal contribution)

Weifeng Hu, and Rui Yang. “Predicting the success of kickstarter projects in the US at launch time”. In Intelligent Systems and Applications (Cham, 2020), Springer International Publishing, pp. 497–506.

ABSTRACT OF THE THESIS

A Data-Driven Study of Cross-Cultural Social Impressions on Faces

by

Weifeng Hu

Master of Science in Computer Science

University of California San Diego, 2020

Professor Gary Cottrell, Chair

Facial impressions play a crucial role in real life, affecting decisions from dating choices to electoral outcomes. Globalization has made it increasingly important to understand how impressions are formed across different cultures. In this thesis, we conduct a large scale data-driven study on cross-culture social impressions of faces. We start by collecting impression ratings of Chinese Asians and U.S. Caucasians on Chinese and Caucasian faces, for 18 social traits related to warmth, attractive-youth, competence and sexual dimorphism. By analyzing the collected data, we observe some widely-agreed upon cultural universals in how high-level facial features relate to impressions. On the other hand, we also find the following cultural differences: (a) Asians give overall lower positive impression ratings to faces compared to Caucasians. (b) raters from both cultures agree more on warmth-related traits, but (c) less on competence-related traits. Furthermore, we introduce a simple and interpretable model, CultureNet, to predict the social impression ratings of a given facial image on specific social traits. The model leverages on the correlation structure of social impression traits and uses them as an attention map that activates different areas of image features. We find CultureNet is able to outperform alternative models with a small number of parameters. Moreover, the trait embedding and activation visualization generated by CultureNet allow us to interpret how the predictions are made. Our work provides a novel method of understanding what features drive facial impressions and allows us to compare perceptions across different cultures.

Chapter 1

Introduction

Although we are told not to judge a book by its cover, we nonetheless do it frequently when we see people for the first time. At the first sight of a new person, our brain automatically forms impressions of them – how trustworthy are they? how kind? what is their social status? Even if these spontaneously formed social impressions are not objectively true [22] (consider the case of Ted Bundy!), they nevertheless affect important aspects of our lives including interpersonal relationships, hiring and financial decisions [25], even legal judgments [36] and electoral outcomes [33, 34].

Regardless of their dubious accuracy, people have fairly high agreement in the facial impressions they form [8]. This agreement is also reflected in the image-level facial features that drive impression formation, such as the apparent age, gender, race and expressions of the face [6, 1, 37]. This agreement also arises in the correlation structure among the impressions of different traits, that seem to fall along three factors: warmth, competence and youthful-attractiveness [34, 31].

Despite these universal aspects of facial impressions, the impressions we form are also influenced by the cultural background of the viewer [34]. This should be no surprise. Research suggests that culture even shapes visual perception [21], and it certainly shapes our social norms, expectations, and values. For instance, East Asians have been characterized as being more collective and holistic, whereas Westerners have been more individualistic and analytic [14, 23]; perhaps this would make friendlier looking people seem more capable to Asian viewers. Moreover, culture also influences our eye movements when we look at faces [4], which may mean that different facial features will be more salient to viewers from different cultures. Altogether, cultural differences in facial impressions seem quite plausible, and their social importance may be increasingly large, given the preponderance of face-to-face international interactions over video conferencing and social media.

Previous studies of cross-cultural facial impressions have identified similarities and differences in a number of individual traits such as attractiveness [5] and intelligence [18]. Yet most prior studies used a small set of strictly controlled face stimuli, limiting the generalizability to everyday face photos with real-world variation. Furthermore, prior studies explored one trait at a time with different face stimuli, compromising any across-trait comparisons in cultural agreement levels. Bridging this gap requires large-scale cross-cultural studies of many traits using a large set of real-world facial images.

More recent work suggests some difference between Chinese and British raters along three previously identified impression dimensions [31]. While this research indicates that there is substantial cross-cultural agreement in warmth and attractive dimension, and the study uses visualization to illustrate the differences in capability dimension, this approach has not attempted to quantify how facial features drive the cross-cultural differences in impression formation, therefore left the mediating mechanisms unaddressed.

Here in this study, we endeavor to understand the cross-cultural universals and idiosyncrasies of facial impressions systematically, and to illustrate the missing link between mediating factors, i.e, facial features, and cross-cultural differences, i.e. how different cultural groups use the same set of facial features differently to form impressions. To do so, we compare how Chinese Asians and American Caucasians (henceforth, Asians and Caucasians, with the country understood) form impressions of 18 traits for each of thousands of real-world Asian and Caucasian face images. As shown in Table 1.1, we consider 18 social impression traits that cover three major categories: (1) warmth related traits, such as warm, happy, friendly, trustworthy, extroverted, humble, calm and kind; (2) physical appearance appraised traits, such as attractive and healthy; (3) capability related traits, such as capable, diligent, high-social status, intelligent, powerful, responsible and successful. As part of the reality check of our data, we also add (4) sexual dimorphism trait - masculine.

Our study provides insights into the mediating factors of cross-cultural perception of faces and suggests directions to future researchers of the important high level facial features related to first impression formation. It demonstrates different potential stereotypes and bias towards certain face features, whose cultural roots are open for future work to investigate.

Table 1.1: The 18 social impression traits in the dataset divided by their major category

Warmth	Capability	Attractive-Youth	Sexual Dimorphism*
calm	capable	attractive	masculine
extroverted	diligent	healthy	
friendly	high-social-status		
happy	intelligent		
humble	powerful		
kind	responsible		
trustworthy	successful		
warm			

* Reality check of the data (not one of the three major categories)

This thesis contains three main parts of the cross-culture study: (1) large-scale data collection (2) a cross-cultural comparison using data analysis (3) modeling and visualizing the cross-culture ratings using neural networks. We will discuss each in details in the following chapters.

Chapter 1 and 3, in full, is currently being prepared for submission for publication of the material. Song, Amanda; Hu, Weifeng; Yadav, Pratap Devendra; Wen, Fangfang; Zuo, Bin; Cottrell Garrison; Vul, Ed. The dissertation/thesis author was the primary investigator and author of this material.

Chapter 2

Data Collection Method

2.1 Face Images

In order to evaluate cross-cultural differences in rating social impressions of faces, we performed a large-scale data collection with Caucasian participants in the U.S. and Asian participants in China. We used 1,836 Caucasian faces from the US 10K Adult Database [3]. For Asian faces, we followed a procedure similar to [3]: We gathered the most frequently used Chinese first names and last names for both genders, and then used the combination of first and last names (in Chinese characters) as the keywords to search images using the Microsoft Bing Image search engine. After the original images were downloaded, we ran a face detector [17] to crop the face region from the image, and removed the images if they met one of the conditions: (1) with a face region resolution smaller than 200 px × 200 px; (2) celebrities (to the best of our knowledge); (3) where at least more than half of the face was occluded; or (4) an infant. After preprocessing, we kept 1,738 Asian faces. Figure 2.1 shows a few examples of the Caucasian and Asian face stimuli.

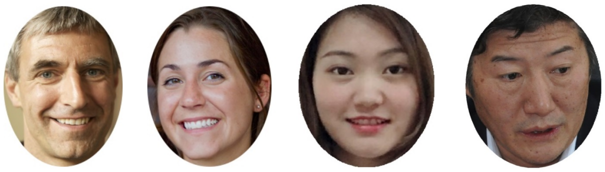


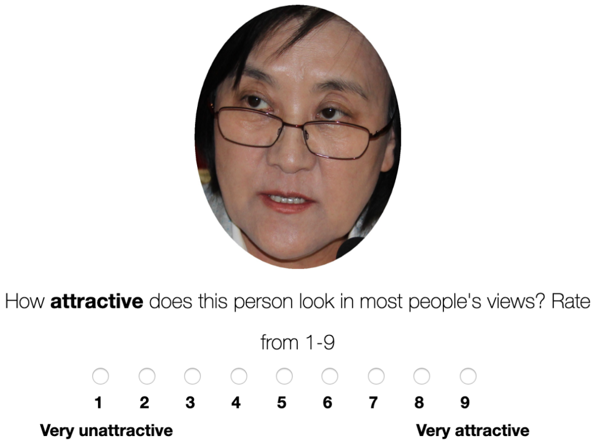
Figure 2.1: Examples of Caucasian and Asian face stimuli.

2.2 Social Impression Traits

We used 18 social impression traits that align with the three key dimensions commonly found in prior research on first impressions of faces [31, 34], plus one sexual dimorphism trait as a reality check: (1) warmth/approachability related traits: friendly, happy, kind, trustworthy, extroverted, humble and warm; (2) attractive/youthful and physical appearance related traits: attractive and healthy; (3) competence related ones: calm, capable, diligent, (of) high social status, intelligent, powerful, responsible and successful; (4) sexual dimorphism related trait: masculine.

2.3 Participants' Task

During the online data collection experiment, participants are asked to indicate their first impression of an image on a specific trait by providing a rating on a scale of 1-9, as shown in Figure 2.2. The data collection experiment contains two parts: First we perform a prescreen task in which 20 randomly selected unique faces of the same ethnicity and a randomly-selected social trait to rate. The 20 faces were presented, shuffled and presented again, which results in a 40-image sequence. If a participant's reliability was significantly above zero (as measured by Spearman rank correlation of test/retest ratings), and he/she used at least three different scales from the 9 point scale, the participant was considered to have passed the "prescreeen test". Otherwise, the participant's data is not used. For each rater ethnicity group, we collected at least 10 ratings per image-trait combination. To achieve this experimental procedure, we built a website that collects these rating data and saves them to an AWS Relational Database.



How **attractive** does this person look in most people's views? Rate from 1-9

○ ○ ○ ○ ○ ○ ○ ○ ○

1 2 3 4 5 6 7 8 9

Very unattractive **Very attractive**

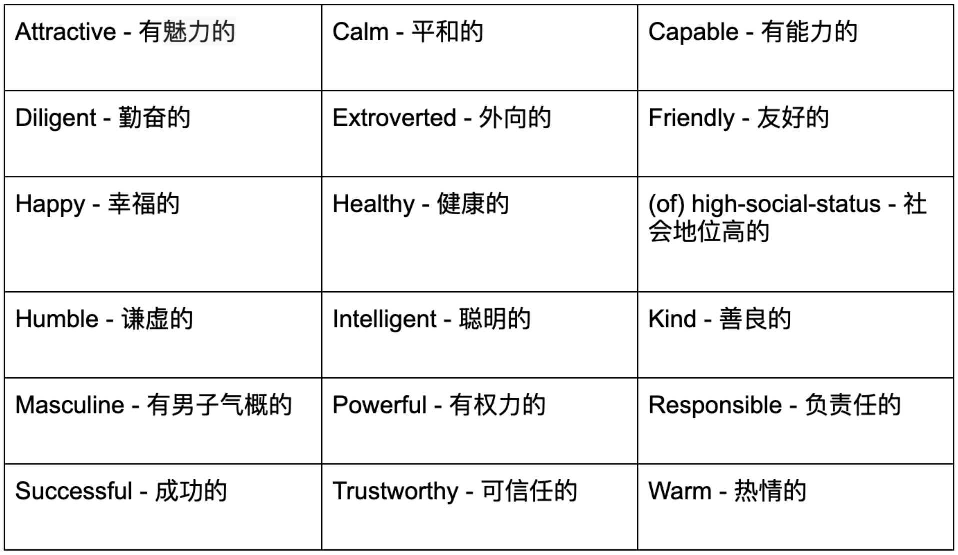
Figure 2.2: First impression rating task page.

2.4 U.S. Data Collection

We performed data collection on U.S. Caucasian participants by sending out the links of our website to Amazon Mechanical Turk workers via TurkPrime [19]. we performed a total of four prescreen rounds. A total of more than 600 subjects participated in our studies. After filtering out data with negative self-correlation and non-Caucasian participants through a survey, we have a subject pool of 428 U.S. Caucasian participants. We then re-invited those participants repeatedly to our main task. We pay each participants 0.25 dollars for prescreen tasks and 0.75 dollars for each 100 images they rate.

2.5 Chinese Data Collection

Since there is no exact equivalent version of Amazon Mechanical Turk platform in China, we recruited Chinese participants via the data100 website (https://www.data100.com.cn) as well as online volunteer sourcing. The task instructions and all traits were translated into simplified Chinese and then back-translated into English to ensure that the Asian participants were rating the same social traits as the Caucasians. Figure 2.3 shows the exact words used in English and Chinese for social impression traits. Due to the limitation of the recruiting platform, we were



Attractive - 有魅力的	Calm - 平和的	Capable - 有能力的
Diligent - 勤奋的	Extroverted - 外向的	Friendly - 友好的
Happy - 幸福的	Healthy - 健康的	(of) high-social-status - 社会地位高的
Humble - 谦虚的	Intelligent - 聪明的	Kind - 善良的
Masculine - 有男子气概的	Powerful - 有权力的	Responsible - 负责任的
Successful - 成功的	Trustworthy - 可信任的	Warm - 热情的

Figure 2.3: The translation of social impression traits from English to Chinese

unable to ask Chinese participants to take multiple main tasks. Therefore we integrated the prescreen process into the main task. Each Chinese participant would see a total of 100 images. The first 40 facial images were used to screen the subjects using the same method as for the Caucasian participants. If a participant passed the prescreen, she/he would continue to see the remaining 60 unique faces. If they didn't pass the prescreen, the task terminated after the first 40 faces. The participants were paid in credits from the recruiting company data100.

In total, we recruited 428 Caucasian raters, 254 of them are females and the median age range is 30-39 years old. Due to the data collection platform differences, much more Chinese Asian participants were recruited: 23,304 in total, 14,338 of them are females and the median age range is 20-29 years old.

2.6 Data Reliability

Since we are dealing with human-rating data, an important sanity check would be on the participant's self-consistency. In our experimental design, we incorporated test/retest methodology. For Caucasian participants, they were asked to rate 10 repeated facial images in the main tasks.

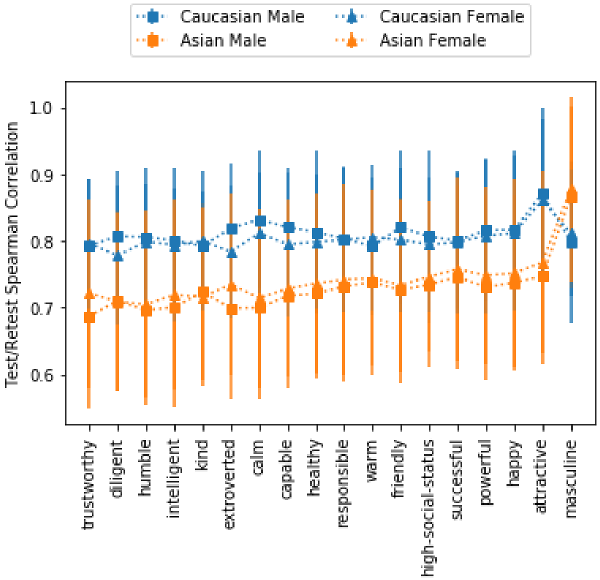


Figure 2.4: The average and standard deviation of self-consistency plots grouped by the ethnicity and gender of the participants. We can see that all groups have high self-consistency values and there is no significant difference in self-consistency among different groups.

For Asian participants, they were asked to complete the 40 repeated prescreen images at first. This design gives us an opportunity to examine the self-consistencies of Caucasian and Asian participants who we chose to include their rating data in our study. We use Spearman correlation as our evaluation metric for consistency. Figure 2.4 shows the average self-consistency plots separated by the ethnicity and gender of the participants. The error bars represent the standard deviation within the ethnicity and gender group. We can see that for all traits, the average self-consistencies for Caucasian and Asian are above 0.65, suggesting a high self-correlation for our participants whose data were recorded in our study. Moreover, the standard deviation suggests that there is no significant difference in self-consistency among different ethnicity and gender groups of the participants. Therefore, we were able to collect a highly reliable rating dataset for the cross-culture facial impressions study.

Chapter 3

A Cross-cultural Comparison of Social Impressions of Faces

3.1 Annotation of High Level Facial Features

In this chapter, we look at high-level facial features of our images. Previous research has shown that multiple facial features, such as the age, gender, ethnicity, and expression of the face are important for facial impressions [6, 1, 37]. To quantitatively examine the influence of facial features on impressions, we need to annotate high level facial descriptors of all the faces in our collected dataset.

We use another large scale dataset, CelebA-HQ[16], which contains 30,000 high quality face images labeled on 40 high-level features. We train a Convolutional Neural Network (CNN) classifier based on [38] by fine-tuning a ResNet-50[12] model pre-trained on ImageNet. Based on the classifier's performance, we remove facial features which had poor classification accuracy. We also removed the highly overlapping facial features. After the pre-filtering, we kept the following 8 high level binary facial descriptors: *gender, ethnicity, smiling, bushy eyebrows, high cheekbones, wearing lipstick, wearing eyeglasses, having beard*. The average classification accuracy for these eight traits is 82.3%. We then took the trained classifier and applied to our own dataset to check whether any of these high-level facial features influences perception of first impressions. In addition to the eight binary facial features, we also obtain an estimate of the face's age using Amazon Rekognition API (https://aws.amazon.com/rekognition/).

3.2 Results

3.2.1 Group Mean Analysis

How does culture and rater ethnicity affect the way people rate faces on average? We compare Caucasian and Asian raters' average rating on each trait respectively, and find that Asians gave lower ratings than Caucasians on 16 out of the 18 traits (the two exceptions being masculine and extroverted; $p < 0.01$). This is consistent with prior observations that European Americans tend to maximize positive feelings and minimize negative ones more than Chinese people do [29].

First, we examined how Caucasian and Asian participants rated faces differently on average for each trait. We divided the participants by ethnicity and subdivided the face images into four demographic groups according to the race and gender of the face. Then, for each face image group, we plot the mean ratings across all Asian raters against all Caucasian raters. The results are shown in Figure 3.1.

We observe that Asian raters give overall lower ratings than Caucasian raters. All of the ratings in Figure 3.1, including happy, are significantly higher for Caucasians over Asians ($p < 0.01$). This trend aligns with prior results arguing that compared to Chinese participants, European Americans tend to emphasize the positive, and downplay the negative [29].

Second, we find that on average, images of Caucasians are rated higher than images of Asians across all traits ($\beta = 0.22$, se $= 0.008$), in particular for warmth related traits ($\beta = 0.41$, se $= 0.014$). However, smiling seemed more common among the Caucasian faces than Asian faces in our pseudo-randomly sampled image set. To correct for this we tagged whether a facial image is smiling using AWS Rekognition (https://aws.amazon.com/rekognition/). We found that 75% of Caucasian images were smiling, while only 31% of Asian images were. Correcting for the effect of smiling reverses the image ethnicity effect, such that warmth related traits are rated lower for Caucasian smiling images than Asian smiling images ($\beta = -0.14$, se = 0.017), and lower for Caucasian non-smiling images than Asian non-smiling images ($\beta = -0.56$, se = 0.019). Table 3.1 shows the dramatic disparities in smiling rates and the reversal of the Caucasian advantage when smiling is controlled. This pattern of results is suggestive of raters implicitly correcting for the different base rate of smiles among Asian and Caucasian faces; thus making a smile more diagnostic for Asian faces, and a lack of smile more diagnostic for Caucasian faces. Regardless of the specific reason, the direction and magnitude of the mean difference in ratings for Caucasian images appears to be driven entirely by the preponderance of smiles in Caucasian images, not due to differences in how Asians and Caucasians are perceived.

Table 3.1: Average ratings across all warmth related traits when separating images by ethnicity and whether they are smiling.

		Asian	**Caucasian**	**%**
		Raters	**Raters**	
Non-smiling	Asian	4.56	4.34	69%
Non-smiling	Caucasian	4.13	3.65	29%
Smiling	Asian	5.52	6.72	31%
Smiling	Caucasian	5.60	6.35	71%

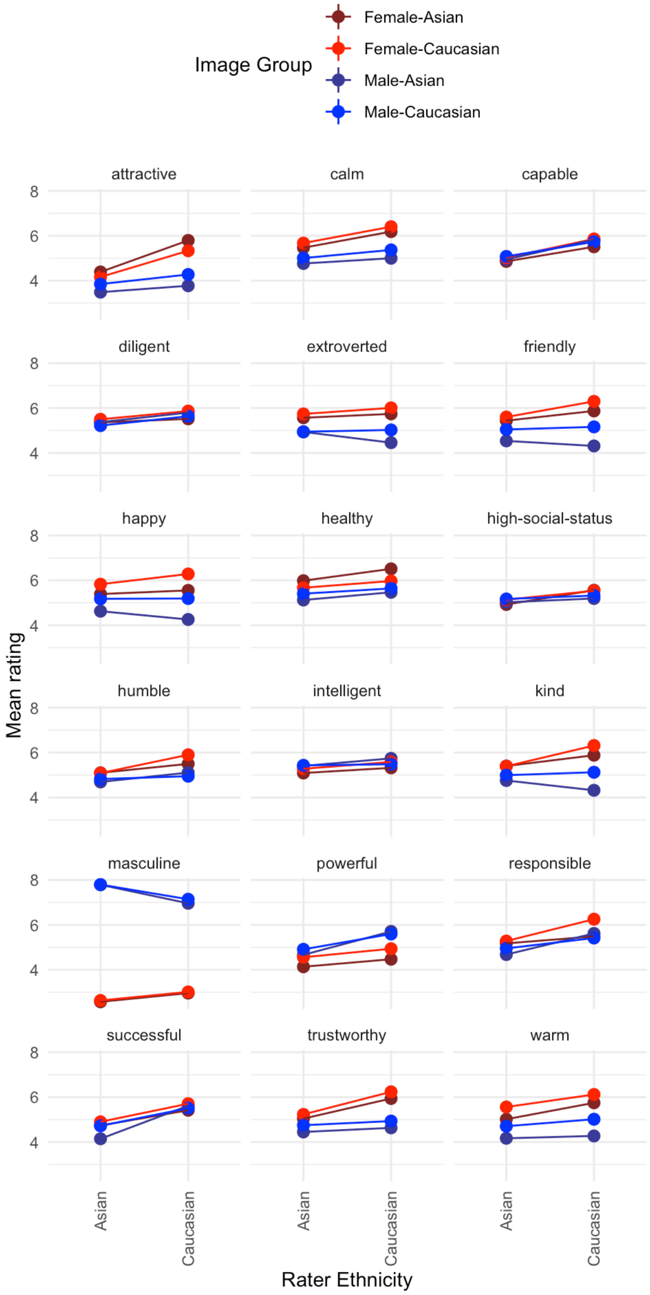


Figure 3.1: For each trait, we split the images based on the gender and ethnicity of the face, and assessed Caucasian and Asians raters' mean ratings and standard errors for the four image groups. Overall, Caucasian raters give higher mean ratings on faces and Caucasian faces in general receive higher ratings.

3.2.2 Consistency Analysis

Impressions of faces are subjective in nature and everyone is their own judge and may hold a unique opinion. Yet, one property that makes facial impressions interesting is the consensus people share to a certain degree. In other words, universality and idiosyncrasies are two sides of the same coin for first impressions. Universality, the agreement people share on a certain impression, may reflect the influence of our common evolutionary history on our perception of impressions; idiosyncrasies, the disagreement people bear on a certain impression, on the other hand, may reflect the influence of our unique environment, including culture, personal history, and our bias on our impression formation.

For each of the 18 impression traits at hand, we can measure the consensus degree at three levels: (1) individual level: how consistent and reliable an individual is when evaluating a certain impression trait; (2) intra-ethnicity level: how much agreement people from the same ethnicity group have (3) inter-ethnicity level: how much agreement Asian raters have with Caucasian raters. These three levels serve as anchoring points to each other, and offer us a rich glance of the convergence and divergence of opinions of various impression traits.

Individual Consistency

Since giving impression ratings is a highly subjective task, we conducted a sanity check to ensure the quality of our data. In particular, we computed the test/retest Spearman correlation [39] on the repeated trials during the screening phase for each rater. Given we have 20 repeated trials, the threshold for significantly above zero ($p < 0.05$) is 0.38. The actual average Spearman correlation is above 0.65 for both rater ethnicities, giving reassurance that our participants were self-consistent and reliable.

Intra-group Consistency

After confirming the individual level consistency, we next examine the agreement level within each rater ethnicity for each trait. We further split each ethnicity group by gender to find out whether there is any potential gender difference in agreement levels.

We used one-way intraclass correlation coefficient (ICC) to measure the agreement level within a group by evaluating the ratio of the variance of item random effects to the overall rating variance. Figure 3.2 shows the ICCs of each trait for each demographic participant group ranked by overall average ICC among all traits. Asian raters have a lower ICC than Caucasian raters; this lower group-level consistency among Asian raters may reflect more diverse opinions about how to evaluate these social traits.

Within the same ethnicity group, there were no statistically significant differences between male and female participants. Similar to previous research [13], we found that there is more agreement for traits representing warmth and appearance-based appraisals (e.g., happy, warm, friendly, kind, attractive), than for competence-related traits (such as diligent, capable, intelligent, and powerful); this effect should not be too surprising as attractiveness, youth, and propensity to smile are much more evident in a picture than traits like diligence.

Inter-group Correlation Analysis

Now we address the third level of consistency, the inter-ethnicity-group consistency, to understand what impressions are perceived more universally and what impressions are perceived differently across cultures. In order to quantify the agreement level Caucasians and Asians have on each impression trait, we use Spearman correlation as a metric, and compare the inter-group correlation levels among 18 impression traits. For each impression trait, we further ask, whether the agreement level differs depending on the face ethnicity, to check if there is an in-group effect.

The results are shown in Figure 3.3. Here, the dots represent the Spearman correlation between Caucasian and Asian participants. All correlations are statistically significant. Since

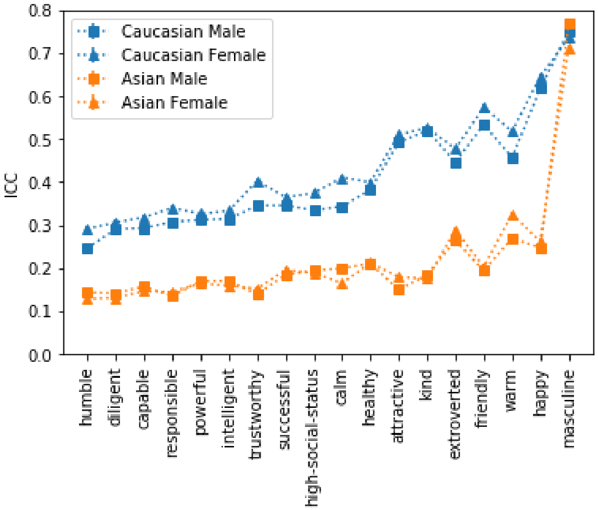


Figure 3.2: ICC raters grouped by ethnicity (Asian/Caucasian) and gender (male/female). Traits are sorted from low to high based on average ICC. The warmth related traits also have high agreements, compared to competence related traits. Overall, Caucasians have higher ICCs than Asian raters, regardless of the rater gender.

we are more interested in figuring out where Asian and Caucasian disagree with each other, we take a closer look at the three least agreed-upon traits: responsible, humble and successful. In particular, Asian and Caucasian raters disagree more on Asian faces regarding these three traits.

To qualitatively examine the differences in the ratings on responsible, successful, and humble, we rank facial images based on how differently they are rated on average by Caucasian and Asian raters, then divided by the sum of standard deviation within each rater ethnicity group. i.e., $\frac{\mu_{Asian}-\mu_{Caucasian}}{\sqrt{\sigma^2_{Asian}+\sigma^2_{Caucasian}}}$. In this way, we can see the faces that are rated higher by Asians than by Caucasians, and vice versa on several traits, as shown in the Figure 3.4. For each trait, the nine images on the left are rated higher by Caucasian participants and the nine images on the right are rated higher by Asian participants, after correction on variance. For the purpose of face identity protection, we morphed several faces together to form each exemplar we show.

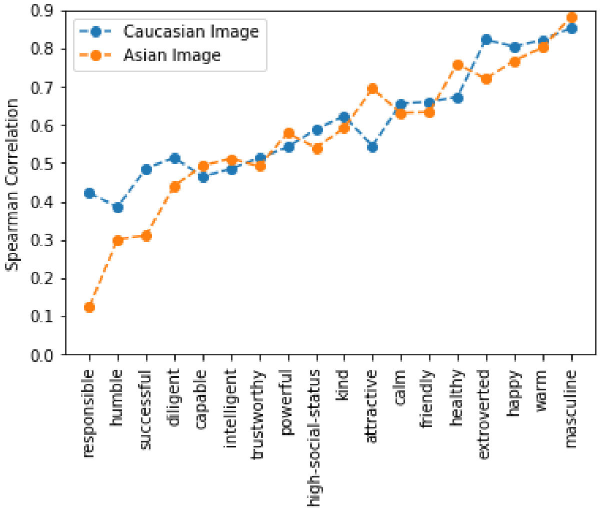


Figure 3.3: We split the images by the ethnicity, and plot the Spearman correlation to assess how Caucasian and Asian raters agree with each other on various traits, as well as the difference in agreement levels for Asian and Caucasian faces. On the left side, we can see that they have overall low agreement on responsible, humble and successful. In particular they disagree more on Asian faces.

3.2.3 Age-based Analysis

Based on visually inspecting the extreme face examples, we see that for all the three traits, faces rated higher by Caucasian seem to be of older age than those rated higher by Asians. We hypothesize that age might play a role in Caucasian and Asian raters' different facial impressions. To test if this hypothesis holds true beyond the few examples we see, we label the perceived age of all the faces in our database, using an age-tagging API by Amazon Rekognition (https://aws.amazon.com/rekognition/). Each face is given an age estimation range, and we take the average of the range as the approximate age of the face.

We divide the face images into different age ranges (in five-year intervals), from below 20, all the way to above 55, and then we divide the raters and faces by ethnicity and plot the average rating as a function of age for each rater ethnicity-image ethnicity combination for the three impression traits we examined above. As we can see from Figure 3.5, the overall trends are Caucasian raters give higher ratings to more senior people whereas Asian raters give no higher,

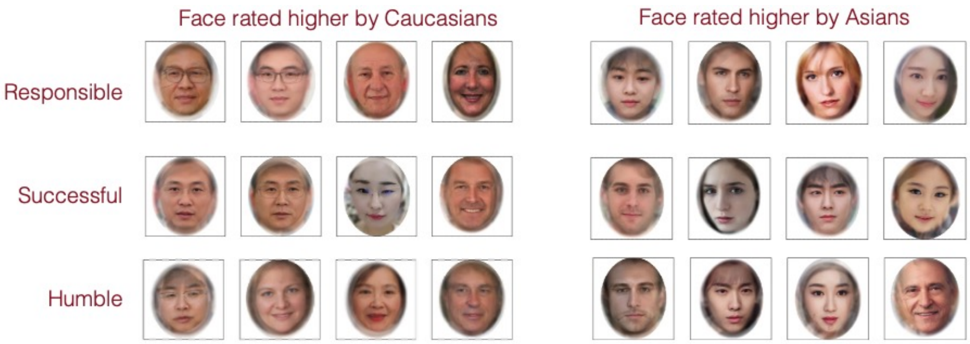


Figure 3.4: Images (in morphed format for privacy protection) that are rated most differently by Caucasians and Asians in responsible, successful and humble(from top to bottom). Images on the left side are rated higher by Caucasians than Asians, whereas images on the right side are rated higher by Asians than Caucasians.

if not lower ratings to more senior people. This trend holds true for all three traits: successful, humble and responsible, although with varying degrees of difference.

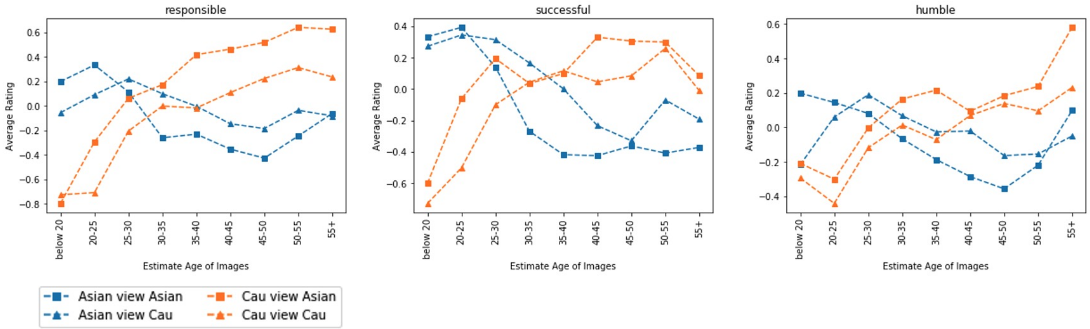


Figure 3.5: Faces are tagged into different age ranges, in five- year intervals. Blue lines are Asian raters average ratings on faces and orange lines are Caucasian raters' average ratings. Images are split by the ethnicity of the face as well: square indicates ratings on Asian faces and triangle indicates ratings on Caucasian faces.

To further quantify the degree of differences in Caucasian vs Asian raters' responses to age, we fit a linear regression model for Asian raters and Caucasian raters respectively for each social trait, using the age of images to predict the average rating of images. We divided the data by the ethnicity of the raters, and we split images into one of the nine categories based on estimated age from below 20 to above 55. For each rater ethnicity group, we took the average of the z-scored ratings for each rating as the final ratings. A linear regression model is then used to fit the final ratings vs estimated image ages (continuous). Lastly, we visualize the coefficients of the linear model using a heatmap. We plot the slope of each model on each trait for Caucasian and Asian raters in Figure 3.6. The first column shows the overall common effect, the second and the third columns show the deviation from the effect for the different rater ethnicity. As we can see from the figure, the overall effect are larger than the ethnicity specific effects, but there are considerable ethnicity-specific variations, most notably in attractive, responsible, successful and humble.

From this figure, we can see that for responsible, humble and successful, Asian participants and Caucasian participants have coefficients of opposite signs. More specifically, being senior makes one look more responsible and successful in Caucasian people's eyes. More senior people look more humble and diligent for Caucasians, but more of high social status and less successful for Asian people. This confirms our hypothesis that age is judged differently when forming impressions.

3.2.4 Lasso Regression Model on Social Impression

Lasso regression model

In order to quantitatively understand to what extent these high level facial features contribute to the perception of social impressions, we train a Lasso regression model that uses the high level facial features to explain and predict the social impression traits.

One of the benefits to use a Lasso model is that its L1 regularization will drive coefficients of unrelated facial features to zero, therefore giving us a concise high level picture with the remaining non-zero coefficients.

The Lasso models are trained on our cross-cultural dataset of Caucasian and Asian images, with each image rated on 18 social impressions and nine high level facial features.

We train a separate model for social impression ratings from Caucasian raters and Asian

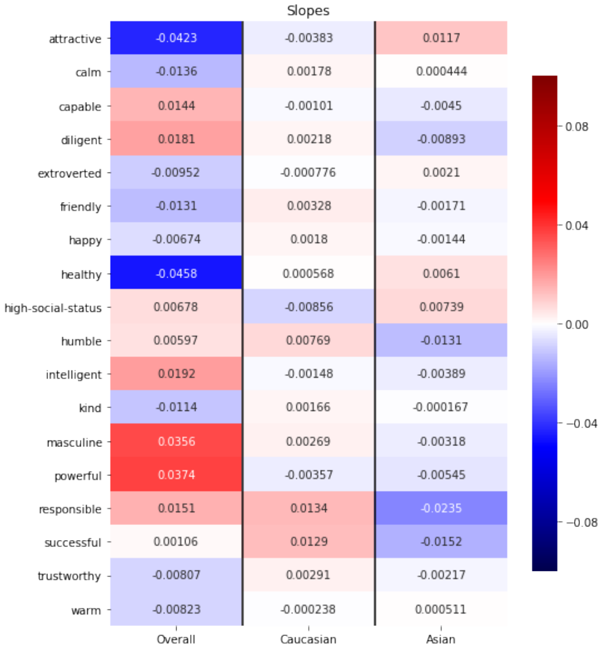


Figure 3.6: Slope of our linear regression model for each trait, when dividing the rating data by rater ethnicity and images into one of the nine age categories from below 20 to 55+. The first column represents the overall trend with both Caucasian and Asian data. The values in the second and third columns are the magnitude of deviations from the overall trend. The scaling color is encoded by magnitude of the coefficient (blue for negative values, red for positive values).

raters, respectively. With the two trained models, we can visualize the learned coefficients to identify most relevant facial features for every social impression trait, and compare them across rater ethnicities.

Results and discussion

We evaluate our Lasso regression models using coefficient of determination (R^2) and the Spearman's rank correlation, which is calculated between the predicted and ground truth human social trait ratings.

For Caucasian raters, the Lasso model achieves $R^2 = 0.41$ and Spearman's correlation = 0.62. For Asian raters, the Lasso model achieves $R^2 = 0.25$ and Spearman's correlation = 0.46. We observe that the higher noise in ratings from Asian subjects causes a drop in R^2 and

Spearman's correlation. We visualize the coefficients from Lasso model trained on Caucasian and Asian raters in Figure 3.7 and Figure 3.8 respectively. In both figures, we sort the nine facial features from left to right based on their average absolute magnitude, so the more important features are on the left. Age and smiling are the two most important factors whereas features like having bushy eyebrows or beard are relatively less important. On the 18 social impression traits, we sort them based on the four broad categories: warmth-related ones, capability-related ones, attractive-youth and masculine. We find a similar categorization of social traits upon performing K-means clustering on the lasso model's coefficients, with K=4 using Scikit-Learn [24].

First, we look at each facial feature's overall effects on each broad category, e.g. whether age has a positive, negative or neutral effects on warmth-related perception. We compute the average coefficient of one specific facial feature on one broad category from one rater ethnicity model. We define the overall effect positive when it's above 0.1, negative when it's below -0.1, neutral when it's in between. For most of the facial features, the effect is the same for Caucasian raters and Asian raters, there are only two exceptions. The general trends and exceptions are summarized in Table 3.2. Smiling, wearing lipstick are positive factors for warm, capability and attractive-youth related perceptions. Apart from that, for warmth-related impressions, the other positive factors are the face "is Asian", and having high cheekbones. The negative factors are apparent age of the face and the face "is-male". For capability-related impressions, age, is male, and wearing eyeglasses are positive, and having beard is negative. One cultural difference is that a face is Asian has a negative impact on capability-related impressions, but only for Asian observers. For attractive-youth, bushy eyebrows is a positive factor and age plays a negative role. For masculine perception, not surprisingly, is male is positive and wearing lipstick is negative. Age is a positive factor, but only for Caucasian people.

After examining facial features' overall effects on broad impression categories, we then zoom in to identify the culturally different patterns at single impression trait level. For each facial feature - impression trait combination, we have a pair of coefficients, one from the Caucasian people's model and one from Asian people's model. We compare the absolute difference of the two coefficients: if it's above 0.55 (95% percentile of all the coefficients), then we consider this facial feature has a culturally different impact on this impression trait. Based on our threshold criteria, we find Caucasians and Asians form the following impression traits differently: Wearing eyeglasses makes one look more responsible and successful in Caucasian people's eyes. Age decreases attractiveness perception more for Caucasians than for Asian observers. Smiling increases trustworthy perception more for Caucasians than for Asian observers.

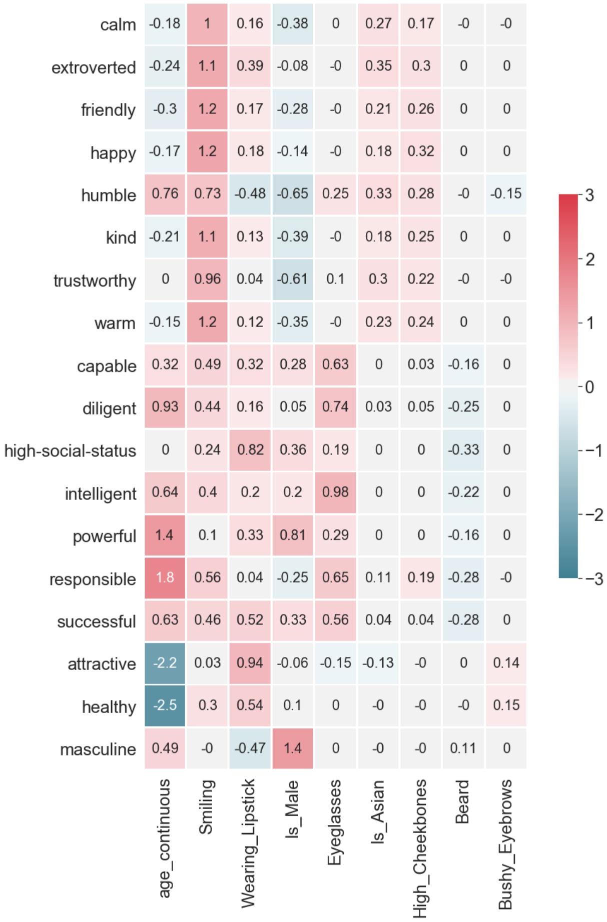
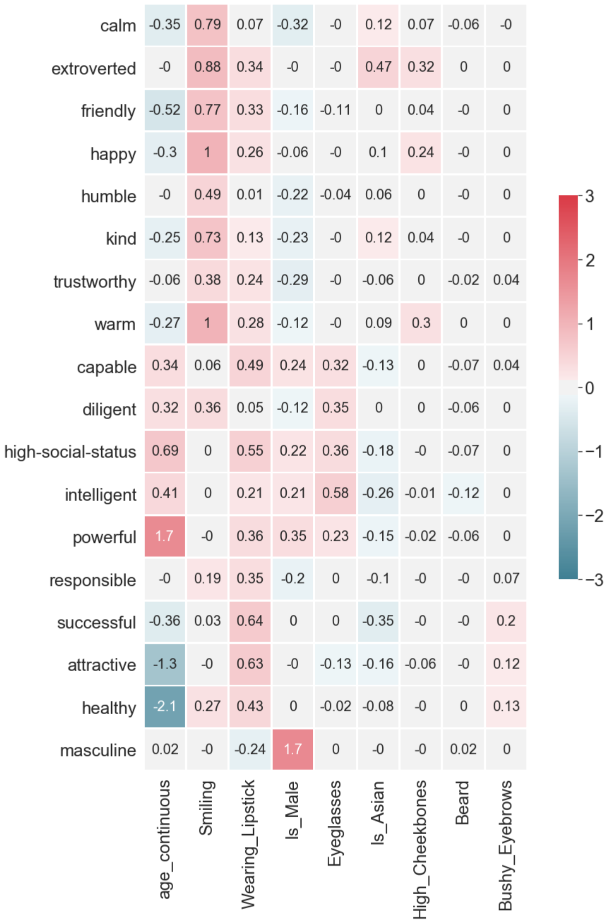


Figure 3.7: Heatmap for coefficients learned by a lasso regression model fit on Caucasian raters' data. The high level facial features (model inputs) and social impression traits (model outputs) are shown along the columns and rows respectively.



	age_continuous	Smiling	Wearing_Lipstick	Is_Male	Eyeglasses	Is_Asian	High_Cheekbones	Beard	Bushy_Eyebrows
calm	-0.35	0.79	0.07	-0.32	-0	0.12	0.07	-0.06	-0
extroverted	-0	0.88	0.34	-0	-0	0.47	0.32	0	0
friendly	-0.52	0.77	0.33	-0.16	-0.11	0	0.04	-0	0
happy	-0.3	1	0.26	-0.06	-0	0.1	0.24	-0	0
humble	-0	0.49	0.01	-0.22	-0.04	0.06	0	-0	0
kind	-0.25	0.73	0.13	-0.23	-0	0.12	0.04	-0	0
trustworthy	-0.06	0.38	0.24	-0.29	-0	-0.06	0	-0.02	0.04
warm	-0.27	1	0.28	-0.12	-0	0.09	0.3	0	0
capable	0.34	0.06	0.49	0.24	0.32	-0.13	0	-0.07	0.04
diligent	0.32	0.36	0.05	-0.12	0.35	0	0	-0.06	0
high-social-status	0.69	0	0.55	0.22	0.36	-0.18	-0	-0.07	0
intelligent	0.41	0	0.21	0.21	0.58	-0.26	-0.01	-0.12	0
powerful	1.7	-0	0.36	0.35	0.23	-0.15	-0.02	-0.06	0
responsible	-0	0.19	0.35	-0.2	0	-0.1	-0	-0	0.07
successful	-0.36	0.03	0.64	0	0	-0.35	-0	-0	0.2
attractive	-1.3	-0	0.63	-0	-0.13	-0.16	-0.06	-0	0.12
healthy	-2.1	0.27	0.43	0	-0.02	-0.08	-0	0	0.13
masculine	0.02	-0	-0.24	1.7	0	-0	-0	0.02	0

Figure 3.8: Heatmap for coefficients learned by a lasso regression model fit on Asian raters' data. The high level facial features (model inputs) and social impression traits (model outputs) are shown along the columns and rows respectively.

3.3 General Discussion:

We conduct a large-scale cross-cultural study of facial impressions, investigate the mediating factors underlying impression formations, and the culturally universals and idiosyncrasies regarding how Caucasians and Asian use these facial cues to form impressions of Caucasian and Asian faces.

First, we find that there is a significant difference between Caucasian and Asian participants regarding their group agreement levels, which suggests that Caucasian participants tend to judge most traits similarly, whereas for Asian participants there are diverse opinions on most

Table 3.2: Facial features' overall effects on broad impression categories

Category	Positive factors	Negative factors	Cultural differences
Warmth	Smiling, wearing lipstick is Asian, high cheekbones	Age, is male	
Capability	Smiling, wearing lipstick Age, is male, eyeglasses	Beard	Is Asian is negative for capability impression on Asian raters only
Attractive-youth	Smiling, wearing lipstick bushy eyebrows	Age	
Masculine	Is male	Wearing lipstick	Age is positive on Caucasian raters only

of the traits in facial images. Although we have already control the sample size with the ICC measure, the difference may or may not have a cultural root since Asian and Caucasian raters are recruited via a slightly different procedures due to practical difficulty in matching the recruitment methods strictly.

Overall, we find that Caucasians give higher ratings than Asian participants on almost all the positive traits. Caucasian faces in general receive higher ratings on warmth related traits, likely due to the fact that there are more smiling faces in the Caucasian face sample pool. Once conditioned on smiling or not, Asian faces receive higher ratings than Caucasian ones. Among the 18 impression traits, people disagree more on competence related ones and agree more on warmth related ones. Among all impression traits, people disagree most on responsible, humble and successful, and they disagree more on Asian faces than Caucasian faces.

By visualizing the faces that are rated most differently by Caucasians and Asian in these three traits, we spot a pattern that faces rated higher by Caucasians are older than faces rated higher by Asians in these three traits. To test if this trend holds true across the whole dataset, we give an age label to every face in our dataset using an API from Amazon. We plot the average ratings on faces as a function of age for Caucasian raters as well as for Asian raters, and find that the trend we observe in the few face examples indeed hold true for the whole dataset. A regression model further gives a quantitative measure of the slope and validate that Asians and Caucasians respond to age differently when forming impressions of responsible, humble and successful.

We extend our attention to more high level facial features, and probe the relationship between more facial features and social impressions. We take advantage of a fully labeled dataset to train classifiers that automatically label our own face stimuli with the high level features of interests. We focus our attention to nine facial features: age, gender, ethnicity of the face, smiling or not, bushy eyebrows, high cheekbones, wearing lipstick, wearing eyeglasses, and having beard. We train two sets of Lasso models on Asian and Caucasian people's rating data, respectively, and by comparing the coefficients from the two races' models, identify the cultural similarity and difference. Regarding each facial feature's influence on the broad impression categories, the influence is the same for both cultures in most cases, the only two exceptions are: Asian faces have a counter-productive effect on capability-related impressions only for Asian raters; age has a positive influence on masculine perception only for Caucasian raters. The following effects are culturally universal: smiling and wearing lipstick have positive effects on warmth, capability and attractive-youth related perception. The detailed quantitative contribution of each facial feature on each impression can be read directly from the coefficient table, and we find out the following culturally different patterns: age is used very differently by observers from two cultures: more senior people look humbler and more diligent for Caucasians, more of higher social status yet less successful in Asian observers' eyes. Age also decreases attractiveness perception more for Caucasians than for Asians. Caucasian observers value smiling more in trustworthy perception. Lastly, wearing eyeglasses and being senior makes one look more successful and responsible, but only in Caucasian people's eyes.

It is worthy further investigate to understand the deeper cultural root behind the different ways the two races use these facial features to form impressions. Previous research suggests that Asian culture has a respect for elderly, how does this cultural tradition connect with the exact impressions Asians have for elder people?

As we mentioned earlier, due to practical constraints, Asian and Caucasian raters are recruited from different channel and their demographics could be better matched in future studies.

Our large-scale cross-cultural dataset also enables computational social scientists to build a more representative, diverse and inclusive algorithm which can predict and modify impression based on Caucasian and Asian people's preferences. Further studies combined with GAN models can establish more explicit causal relationship between facial features and social impression traits, e.g. adding beard directly on images to validate if beard indeed decreases capability impressions, as found in our current study.

Our dataset and statistical analyses provide new perspectives for cross-cultural studies of facial impressions. They not only highlight interesting patterns on how Caucasian and Asian participants use high level facial features similarly and differently to form impressions, but also advance our understanding of the mediating mechanisms underlying social impressions across cultures and provide insights to look for deeper cultural roots to explain people's impression formation patterns.

Chapter 1 and 3, in full, is currently being prepared for submission for publication of the material. Song, Amanda; Hu, Weifeng; Yadav, Pratap Devendra; Wen, Fangfang; Zuo, Bin; Cottrell Garrison; Vul, Ed. The dissertation/thesis author was the primary investigator and author of this material.

Chapter 4

CultureNet: Predicting Social Impression Ratings Using Trait Embedding with Limited Data

4.1 Introduction

Humans quickly form social impressions to judge people they are seeing for the first time. These impressions, although inaccurate, are unfortunately irresistible and influential. For example, when we use the Tinder app, we rely on our own judgment of attractiveness, based mostly on images of faces, to screen possible mates. Previous research has shown that such impressions not only affect our interpersonal relationships, but also hiring decisions [25] and even election outcomes [34]. Despite the subjective nature of forming social impressions, there is a great deal of consensus among raters on a number of traits [34]. Recent studies have found that social impressions fall mainly into three dimensions: warmth, competence and attractive-youth [34, 31, 32]. This suggests that it may be possible to characterize how people use the raw face image to form complex judgments about individuals.

Despite the systematic correlation structure of Western social impressions, recent research has shown that people from different cultural backgrounds are likely to form different impressions of the same social traits [34, 5]. This suggests that different facial features might have different effects on how people of different cultural backgrounds perceive and evaluate faces. For instance, one prior study found that Caucasian observers consistently fixate the triangle formed by the eyes and mouth, whereas Asian observers fixate more on the central region of the face [4].

Recent developments in deep learning have allowed researchers to predict human judgments for a variety of tasks and stimuli, including faces. Models that predict a single social impression trait – most often attractiveness – from face images can achieve high correlations with human judgments [7, 11, 2, 26]. There are also studies that build deep learning models to predict multiple social impression traits [20, 30]. However, there has been scant research into the differences in social impression ratings between different cultures. Part of the reason is that there are few large scale datasets that contain facial images from different cultures. For example, the US 10K Adult faces dataset [3] contains facial images of U.S. subjects only.

Therefore, we used our dataset of social impression ratings on more than $3,000$ facial images with U.S. Caucasian raters and Chinese Asian raters. Using this data, we introduce a novel model called CultureNet. We show that by learning a latent trait embedding, CultureNet can achieve good performance using a relatively small amount of data and a large number of traits. Moreover, CultureNet is able to achieve better performance with a small number of parameters compared to alternative models on this dataset, and the social trait embedding that CultureNet generates captures the systematic relationships among different impression traits. Finally, we show that due to the efficiency of this model, it can characterize cultural differences by locating the regions of the faces that contribute most to the ratings of certain impression traits for raters of different cultural backgrounds.

The main contributions of this chapter are:

- We propose a simple yet powerful architecture that shows good performance in predicting a large number of social impression ratings using a small number of parameters and a relatively small amount of data.

- We leverage the trait embedding generated by our proposed model to further study the relationships between social impression traits.
- We visualize the activation regions to visualize cross-cultural differences in rating certain social impression traits.

4.2 Datasets

We use our facial social impression dataset, containing 1,836 Caucasian images and 1,737 Asian images, rated 35,000 times by Chinese Asian and U.S. Caucasian raters on 18 social impression traits. The impression traits were grouped into four major categories, the first three of which are the underlying dimensions of first impressions found in [34, 5, 32]: (1) warmth: calm, extroverted, friendly, happy, humble, kind, trustworthy and warm; (2) capability: capable, diligent, high-social-status, intelligent, powerful, responsible and successful; (3) attractive-youth: attractive and healthy; and (4) sexual dimorphism: masculine. The final category was added as a sanity check on the data. Each image is rated at least 10 times on each trait. We use the average ratings R_{avg} for each image as the final ratings. Example face images are shown in Figure 4.1.

Figure 4.1: Examples of Caucasian and Asian face images.

We use the average rating for each image-trait combination as the final rating for each rater ethnicity, yielding a total of 3,573 data points for each rater ethnicity. Although this is a huge dataset for cross-cultural perception research, it is a fairly small dataset for training image classification models, highlighting the need for sample-efficient models. We split our dataset, using 80% of the images as training data, 10% as validation data and 10% as testing data. The experimental results are evaluated based on the test data.

4.3 Method

4.3.1 CultureNet

The main idea of CultureNet is that there are a fairly small number of latent facial features that are used in similar ways for traits correlated by the three dimensions listed above. Furthermore, there are likely cultural biases shared by raters of the same culture. If so, we can exploit this latent structure in face features and social impression traits to learn a mapping for a particular trait for a particular rater group using very little data specific to that rater-trait combination. To exploit this structure, we propose the network shown in Figure 4.2. There are three main parts to the network: (1) an image feature extractor E_{im}, (2) a trait feature extractor E_{trait} and (3) a regressor R_e.

The network takes in an input image and a one-hot encoded trait. We learn embeddings for images (from E_{im}) and traits (from E_{trait}) matched in dimensionality and element-wise multiply them to yield a feature vector from which we predict the rating. We use element-wise multiplication so that the trait features are treated as attention maps over the learned image features. Different trait feature vectors highlight different regions of the input image, allowing the model to make different predictions for different traits; however the feature space embedding is shared across traits.

The image is processed through a pretrained image feature extractor network that outputs the image features. We use a MobileNet_v2 [15] pretrained on ImageNet as our image feature extractor and fine tune it on our data. The last classification layer of the MobileNet is removed. For the trait feature extractor, we use two dense layers of 256 nodes and 1,280 nodes respectively.

The one-hot encoded trait is transformed through those two layers to generate the trait features. We then compute the element-wise multiplication of those two feature vectors and the result will go through the regressor for the final rating prediction. The regressor network contains two dense layers of 1,280 nodes and 256 nodes respectively. During training we jointly learn the face-feature embedding, the trait embedding, and the trait prediction regression.

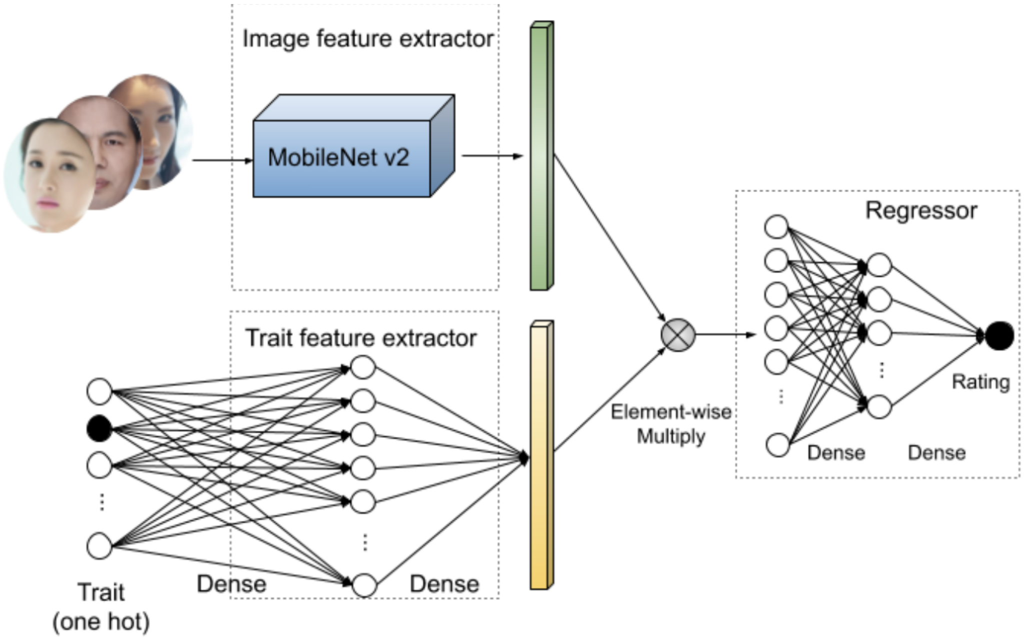


Figure 4.2: Architecture diagram for CultureNet. The green vector is the image feature vector and the yellow vector is the trait feature vector

As this is a regression problem, our loss function $\mathcal{L}$ is mean squared error. When given an input image x_i, we perform a batch update over all traits t with respect to y_{it}. Therefore, the training objective is to minimize the quantity in Equation 4.1:

$$\min \sum_{i=1}^{N} \sum_{t \in \text{traits}} \mathcal{L}(y_{it}, R_e(E_{\text{im}}(x_i) \otimes E_{\text{trait}}(t))) \tag{4.1}$$

4.4 Experiment

We set up our experiment by training a CultureNet model for Caucasian raters (Caucasian model) and a CultureNet model for Asian raters (Asian model). We evaluate our model using two performance metrics, Spearman's ρ, a rank order correlation statistic, and MSE, both comparing the model's predicted values and the average human rating data (R_{avg}) on the test set.

4.4.1 Baselines

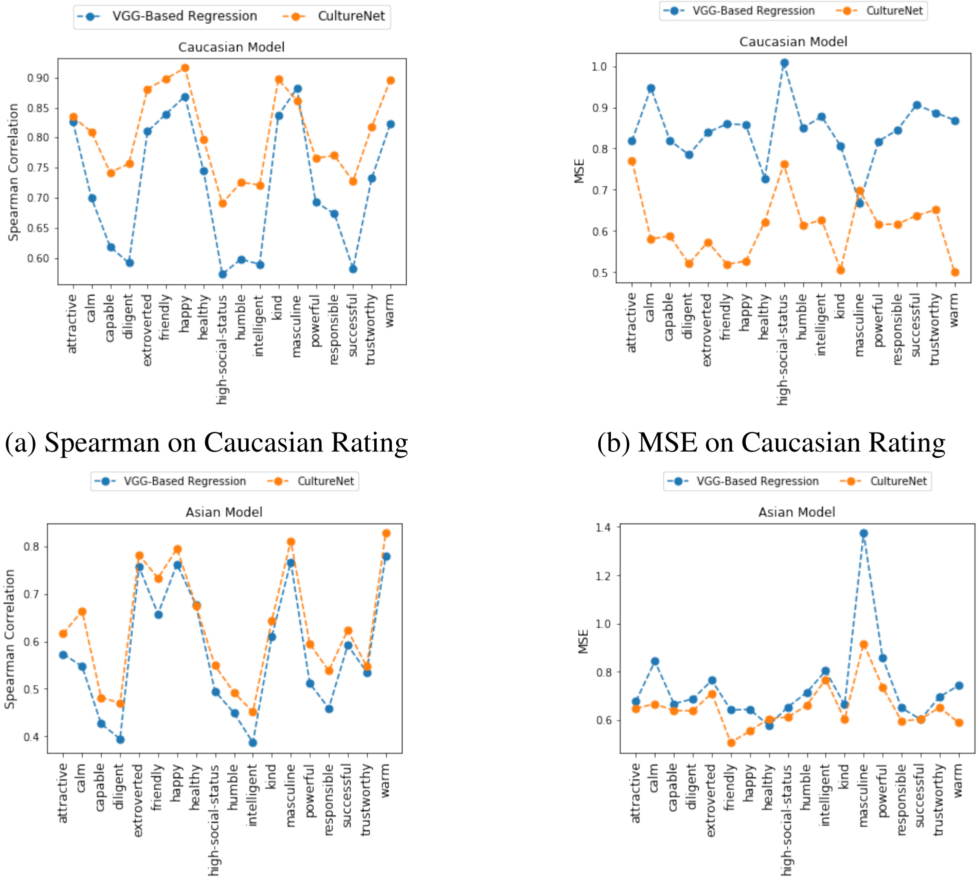
Human Baseline

As a baseline correlation for our model, we calculated the within-rater Spearman correlation and MSE scores. Since the dataset contains ratings for each participant, we use a split half procedure. In particular, we randomly split raters of the same ethnicity into two groups and compute the Spearman correlation and MSE between those two groups. This process is repeated 50 times and the average values are computed to ensure stability. Although this procedure yields numbers on the same scale as our model evaluation, they cannot be directly compared. The human-human correlation can be lower than the model correlation because we split them into two groups and measured the correlation between the groups, whereas the model performance is correlated with the average human ratings. Hence it will be possible for our model to seem to "outperform" the human baseline. Such a result only indicates that the variance between human raters reduces the correlation. Nonetheless, this baseline is useful to evaluate the relative noisiness of different human datasets.

VGG-Based Regression Model

We also implemented the VGG-based regression model from [30]. In this model, a VGG network [28] is used to extract features from the input images. Then PCA is used to reduce the dimensionality of the features (where the number of PCA components is determined by cross-validation separately for each trait) and then regression is used to predict the rating. We compute the Spearman correlation and MSE of the predicted values with the averaged human ratings.

4.4.2 Experimental Results



(a) Spearman on Caucasian Rating

(b) MSE on Caucasian Rating

(c) Spearman on Asian Rating

(d) MSE on Asian Rating

Figure 4.3: Performance of CultureNet and the VGG-based regression model [30] on Caucasian and Asian ratings on each of the 18 traits. The blue line represents the performance of the VGG-based Regression model [30]; and the orange line represents CultureNet.

Figure 4.3 shows the model's performance on Caucasian and Asian rating data for each trait. We would like to achieve high Spearman correlation and low mean squared error. As we can see, CultureNet consistently outperforms the VGG-based regression model [30] over all traits except masculine. Table 4.1 also shows the average Spearman correlation and MSE scores for the human baseline (MSE for humans is calculated in the same manner as the Spearman correlation), the VGG-based Regression model and CultureNet. As expected, we see a clear advantage of CultureNet over VGG-net. Recall that the human baseline is not comparable to the model scores. It is also noteworthy that VGG-based regression model uses more than 160 million parameters, whereas CultureNet model uses only 2.9 million, showing strong performance with a small number of parameters. Of course, with such a small dataset, the VGG-based model could be overfitting.

Table 4.1: Average Spearman correlation and MSE scores of CultureNet model compared to VGG-based regression model with referencing to human.

	Caucasian Spearman↑	**Caucasian MSE↓**	**Asian Spearman↑**	**Asian MSE↓**
Human	0.7681	0.9035	0.4738	1.4611
VGG-based Regression [30]	0.7216	0.8438	0.5767	0.7360
CultureNet	**0.8061**	**0.6071**	**0.6276**	**0.6496**

Discussion

Why is CultureNet able to achieve better performance with fewer parameters? We believe it is due to the use of the trait embedding, which leverages the correlational structure among traits, which in turn helps categorize any one trait better from sparser data. Since most social impression traits can be categorized into three dimensions - warmth, competence and attractive-youth [34, 31, 32], there are indeed correlations among the social traits. The dataset we chose also shows correlation among these traits, as demonstrated in Figure 4.4 which is calculated based on the correlation of ratings in the dataset. As a result, CultureNet takes advantage of this correlation by using similar trait embedding for correlated traits. On the other hand, the VGG-based regression model independently computes the regression for each trait, so it can't use this feature of the data.

To further test this assumption, we first select 5 less correlated traits according to Figure 4.4 - attractive, kind, powerful, successful and masculine. We then randomly select more traits to make a subset of size 8, 12, 15 and 18. We train CultureNet on the Caucasian rating data with these 5 subsets of the 18 social impression traits. We only choose Caucasian data because it is much less noisy than the Asian data, which helps us eliminate noise as a confounding variable. Table 4.2 shows the Spearman correlation and MSE on the test data. We can see that although the Spearman correlation is not affected by different numbers of traits, we achieve better MSE scores when using more traits. When we use 5 least-correlated traits, the mean squared error of CultureNet increases. As we randomly add more traits to the training set, we can see a significant drop in MSE. This explains the advantage of CultureNet with little data but many traits.

Table 4.2: Performance of CultureNet when trained on different size of subset of social traits with Caucasian data. We can see that although Spearman correlation remains unchanged, MSE shows significant improvement with a larger number of traits.

Number of Traits	**5**	**8**	**12**	**15**	**18**
Spearman ↑	0.8002	0.806	0.8203	0.8074	0.8061
MSE ↓	0.7304	0.6741	0.6079	0.6596	0.6071

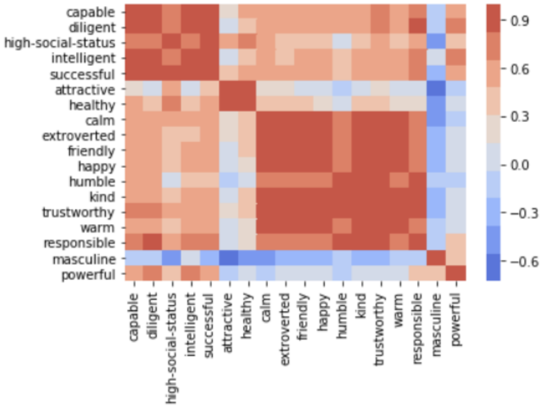


Figure 4.4: Correlation structure of the rating data in our dataset. We can see that many traits are highly correlated with each other.

4.5 Results

4.5.1 Embedding of Trait Features

A notable advantage of CultureNet is that we can take the trait features E_{trait} and obtain the embedding of traits in the dataset. We use the trait embedding as an attention map to activate regions of the images, which results in model outputting an embedding for each trait. We believe the embedding learned in our model represents the latent space of the social impression traits, and thus should resemble the representation obtained from human rating data. We further test this assumption using tSNE visualization [35].

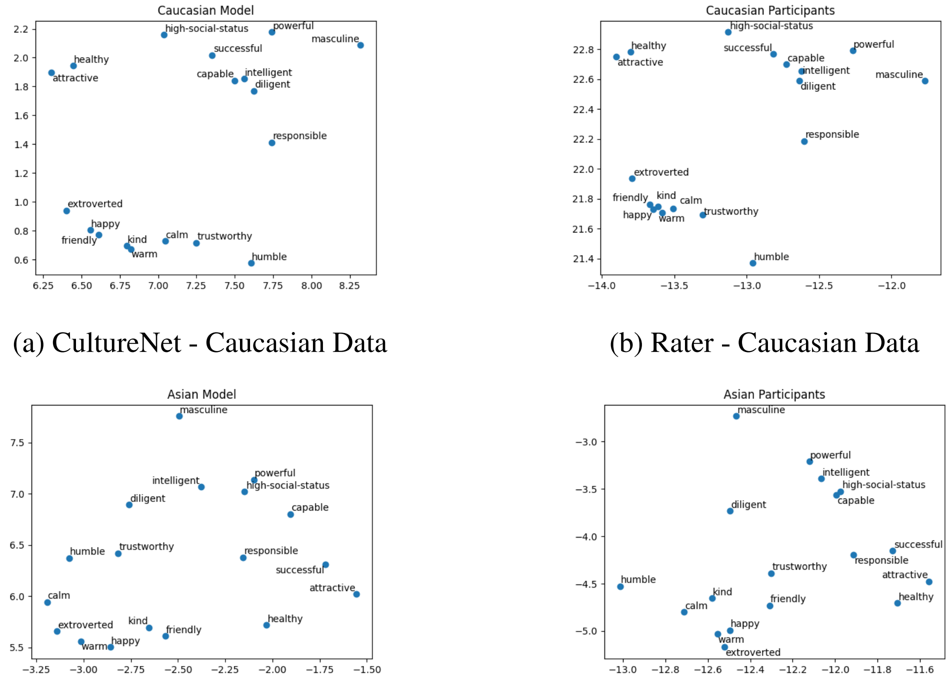
tSNE

We take the one-hot encoded traits and run them through the trait feature extractor E_{trait}, which results in 18 embedding vectors. We then run tSNE with cosine similarity, perplexity of 15 and learning rate of 15 to project data onto 2-d plane. As a reference, we also use raw average ratings across all images to calculate distances between traits and run tSNE with the same hyperparameters.

Figure 4.5 shows the side by side comparison of the tSNE output. Figure 4.5a and Figure 4.5c are the resulting plots using the embedding output by trait feature extractor. Figure 4.5b and Figure 4.5d are the resulting plots using human rating data by computing the cosine distance of average ratings between each pair of traits. As can be seen, the resulting plot shows high similarity of trait features in the embedded space, suggesting the accuracy of the trait embedding output by CultureNet.

4.5.2 Visualization of Activation

One advantage of jointly learning trait embedding and image features is that it allows us to visualize the activation in the input image, which you cannot get from the human rating



(a) CultureNet - Caucasian Data

(b) Rater - Caucasian Data

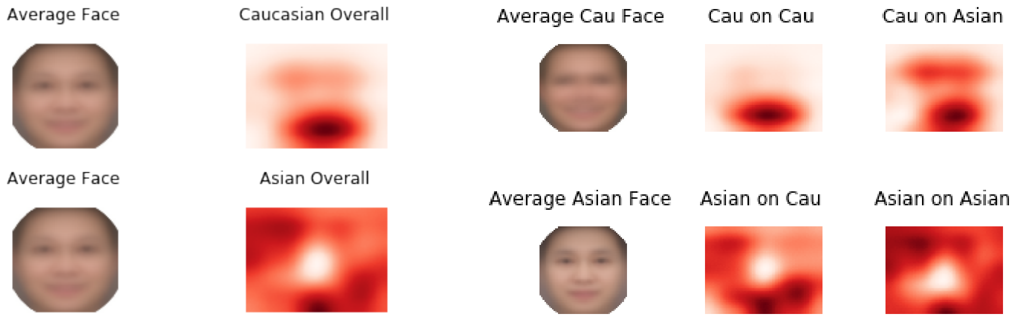
(c) CultureNet - Asian Data

(d) Rater - Asian Data

Figure 4.5: tSNE results of trait embedding from CultureNet trait feature extractor and human rating data.

data alone. We apply GRAD-CAM [27] on the final layer of our image feature extractor in our Caucasian model and Asian model. The resulting activation map provides insight to which image areas the CultureNet model uses when making capturing ratings of different social traits. Since the original GRAD-CAM is for an image classification task, we made a slight modification to their architecture so that it can adapt to our regression task. Instead of having a target class when backpropagating the gradient, we directly backpropagate the model output so that the final visualization will tell us which region contributes positively or negatively to the final prediction.

Figure 4.6 shows the average activation maps over all training samples for our Caucasian model and Asian model. Figure 4.6a is computed using all the images for each model. We can see that for the Caucasian model, the regions that are activated (red) are mainly the eyes and mouth region. For the Asian model on the other hand, the activation tends to be more global around the faces. This is consistent with previous findings that Asian viewers look at faces more globally than



(a) Activation map overall

(b) Activation map by ethnicity of rater × image

Figure 4.6: Activation maps for Caucasian and Asian model on all images as well as on images by their ethnicity group. The red regions indicate that those areas are activated when making predictions.

Caucasian viewers [4]. Figure 4.6b shows the activation map when separating by the ethnicity of images. We can see that for the Caucasian model (first row), the eye region is activated mostly on Asian images. Our guess is that this is likely due to the fact that the only pictures of people wearing glasses in our dataset occur among the Asian images. Thus, eyeglass information is only informative in the set of Asian images. This effect is most salient for Caucasian raters, given their apparently narrow focus, but a similar increased reliance on the eye region for Asian images is also evident for Asian raters.

In Figure 4.7, we choose four social traits for visualization: happy, intelligent, responsible and humble. The first two traits are chosen because there are universal signals (e.g. smiling) that affect the ratings, which we want our model to indicate. The last two traits are chosen because Caucasian and Asian raters differ on them the most. We find that both Caucasian and Asian raters are fairly consistent in their activation maps for Caucasian and Asian images: Pearson product-moment correlations P of the activation maps between the first two rows are high, as are the correlations between the last two rows. Across all 18 traits, the average correlation P is 0.61 ± 0.25 for the Caucasian model and 0.41 ± 0.28 for the Asian model. This suggests that raters of the same ethnicity group tend to process images similarly. For the happy trait, the correlation between activation maps for different image ethnicities is 0.36 for the Caucasian

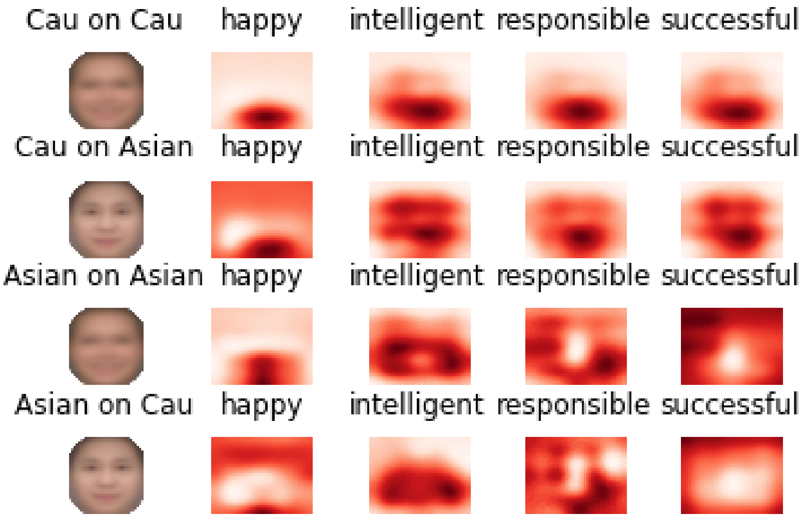


Figure 4.7: Activation map by averaging GRAD-CAM results of all images separated by ethnicity group. The red regions indicate that those areas are activated when making predictions.

model and 0.07 for the Asian model. Despite the low correlation value for the Asian model, we can see that both Caucasian and Asian models activate the mouth region, which suggests they are evaluating whether the face is smiling. On the intelligence trait, the activation maps for Asian and Caucasian images are correlated at $r = 0.80$ for Caucasian raters, and 0.54 for the Asian raters. We can see both activate the mouth and eye regions, consistent with the intuitive, and demonstrable, impression that wearing eye glasses makes people look smarter [9]. For responsible and successful, we can see that Caucasian and Asian models activate very different regions. However, within the two rater ethnicities, activation maps are fairly consistent across image ethnicities (for responsible: $r = 0.8$ for Caucasians raters, and $r = 0.22$ for Asian raters; for successful: $r = 0.89$ for Caucasian raters, and $r = 0.80$ for Asian raters).

Overall, our finding is consistent with previous work arguing that Asian and and Caucasian viewers focus on systematically different parts of faces [4]. We are planning on performing Bubbles technique [10] in the future to validate our activation maps.

4.6 Conclusions and Future Work

We propose a simple and interpretable architecture called CultureNet that can predict ratings of social impression traits. CultureNet achieves better Spearman correlation and MSE scores with human ratings compared to the previous state of the art model, while using far fewer parameters. We believe the advantage of CultureNet lies in its use of trait embedding to leverage the correlational structure among traits, which helps it learn to categorize any one trait better from sparser data. CultureNet is able to achieve good performance when there are many traits with little training data. Moreover, the joint learning of trait embedding and image features by CultureNet allows the visualization of activation in the image space, which points out regions of the faces that contribute most to the rating of each trait. As a result, this introduces a new quantitative method of studying relationships between social impression traits, as well as enabling comparisons between raters of different cultural backgrounds.

For future work, it would be interesting to use CultureNet as a discriminator for a Generative Adversarial Network so that we can generate or modify images which would receive high ratings on a specific trait.

Chapter 4, in full, has been submitted for publication of the material as it may appear in Advances in Neural Information Processing Systems 33, 2020, Hu, Weifeng; Yadav, Pratap Devendra; Song, Amanda; Vul, Ed.; Cottrell Garrison. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 5

Conclusion

In this thesis, we described a data-driven study of cross-cultural social impressions on faces. We start by conducting a large scale data collection with U.S. Caucasian raters and Chinese Asian raters. The images we used are Caucasian images obtained from [3] and Asian images obtained using the Microsoft Bing search. We also pick 18 social impression traits that span the three dimensions of social impressions [31]. The experiments were conducted in different platforms for Caucasian raters and Asian raters but we managed to keep the procedure as same as possible. We find no significant differences in test/retest reliability between Caucasian and Asian raters, which suggests that our data are highly reliable.

After the data collection process, we used various statistical analysis methods to study the cross-cultural effect on social impressions. We observe some widely-agreed upon cultural universals in how high-level facial features relate to impressions. On the other hand, we also find the following cultural differences: (a) Asians give overall lower positive impression ratings to faces compared to Caucasians. (b) raters from both cultures agree more on warmth-related traits, but (c) less on competence-related traits.

Lastly, we built a neural network model, CultureNet, that predicts rating of a specific trait with an input image. CultureNet is able to use trait embedding to leverage the correlational structure among traits, which helps it learn to categorize any one trait better from sparser data. Moreover, the joint learning of trait embedding and image features allows the visualization of activation maps in the image space, which points out regions of the faces that contribute most to the rating of each trait. We find that those activation maps align with previous findings on how Caucasian and Asian perceive faces [4].

In conclusion, our studies provide new perspective for cross-cultural comparison of social impressions of faces. They advance our understanding of the mediating mechanism underlying social impressions across cultures as well as the computational modeling of rating predictions on social impressions.

Bibliography

[1] ADAMS JR, R. B., HESS, U., AND KLECK, R. E. The intersection of gender-related facial appearance and facial displays of emotion. *Emotion Review 7*, 1 (2015), 5–13.

[2] ALTWAIJRY, H., AND BELONGIE, S. Relative ranking of facial attractiveness. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)* (2013), pp. 117–124.

[3] BAINBRIDGE, W. A., ISOLA, P., AND OLIVA, A. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General 142*, 4 (2013), 1323.

[4] BLAIS, C., JACK, R., SCHEEPERS, C., FISET, D., AND CALDARA, R. Culture shapes how we look at faces. *PloS One 3* (02 2008), e3022.

[5] CUNNINGHAM, M. R., ROBERTS, A. R., BARBEE, A. P., DRUEN, P. B., AND WU, C.-H. "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology 68*, 2 (1995), 261.

[6] EBNER, N. C. Age of face matters: Age-group differences in ratings of young and old faces. *Behavior research methods 40*, 1 (2008), 130–136.

[7] EISENTHAL, Y., DROR, G., AND RUPPIN, E. Facial attractiveness: Beauty and the machine. *Neural Computation 18*, 1 (2006), 119–142. PMID: 16354383.

[8] FALVELLO, V., VINSON, M., FERRARI, C., AND TODOROV, A. The robustness of learning about the trustworthiness of other people. *Social Cognition 33*, 5 (2015), 368–386.

[9] FLEISCHMANN, A., LAMMERS, J., STOKER, J., AND GARRETSEN, H. You can leave your glasses on: Glasses can increase electoral success. *Social Psychology 50* (01 2019), 38–52.

[10] GOSSELIN, F., AND SCHYNS, P. G. Bubbles: a technique to reveal the use of information in recognition tasks. *Vision Research 41*, 17 (2001), 2261 – 2271.

[11] GRAY, D., YU, K., XU, W., AND GONG, Y. Predicting facial beauty without landmarks. In *Computer Vision – ECCV 2010* (Berlin, Heidelberg, 2010), K. Daniilidis, P. Maragos, and N. Paragios, Eds., Springer Berlin Heidelberg, pp. 434–447.

[12] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 770–778.

[13] Hehman, E., Sutherland, C. A., Flake, J. K., and Slepian, M. L. The unique contributions of perceiver and target characteristics in person perception. *Journal of personality and social psychology 113*, 4 (2017), 513.

[14] Hofstede, G. Culture and organizations. *International Studies of Management & Organization 10*, 4 (1980), 15–41.

[15] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

[16] Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *CoRR abs/1710.10196* (2017).

[17] King, D. E. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research 10*, Jul (2009), 1755–1758.

[18] Krys, K., Hansen, K., Xing, C., Szarota, P., and Yang, M.-m. Do only fools smile at strangers? Cultural differences in social perception of intelligence of smiling individuals. *Journal of Cross-Cultural Psychology 45*, 2 (2014), 314–321.

[19] Litman, L., Robinson, J., and Abberbock, T. Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods 49* (04 2016).

[20] McCurrie, M., Beletti, F., Parzianello, L., Westendorp, A., Anthony, S., and Scheirer, W. J. Predicting first impressions with deep learning. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)* (2017), pp. 518–525.

[21] Nisbett, R. E., and Miyamoto, Y. The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences 9*, 10 (2005), 467–473.

[22] Olivola, C. Y., and Todorov, A. Fooled by first impressions? reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology 46*, 2 (2010), 315–324.

[23] Oyserman, D., Coon, H. M., and Kemmelmeier, M. Rethinking individualism and collectivism: evaluation of theoretical assumptions and meta-analyses. *Psychological Bulletin 128*, 1 (2002), 3.

[24] PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[25] REZLESCU, C., DUCHAINE, B., OLIVOLA, C. Y., AND CHATER, N. Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PloS One 7*, 3 (2012), e34293.

[26] ROTHE, R., TIMOFTE, R., AND GOOL, L. V. Some like it hot - visual guidance for preference prediction, 2015.

[27] SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D., AND BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *The IEEE International Conference on Computer Vision (ICCV)* (Oct 2017).

[28] SIMONYAN, K., AND ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition, 2014.

[29] SIMS, T., TSAI, J. L., JIANG, D., WANG, Y., FUNG, H. H., AND ZHANG, X. Wanting to maximize the positive and minimize the negative: Implications for mixed affective experience in american and chinese contexts. *Journal of Personality and Social Psychology 109*, 2 (2015), 292.

[30] SONG, A., LI, L., ATALLA, C., AND COTTRELL, G. Learning to see people like people, 2017.

[31] SUTHERLAND, C. A., LIU, X., ZHANG, L., CHU, Y., OLDMEADOW, J. A., AND YOUNG, A. W. Facial first impressions across culture: Data-driven modeling of chinese and british perceivers' unconstrained facial impressions. *Personality and Social Psychology Bulletin 44*, 4 (2018), 521–537.

[32] SUTHERLAND, C. A., OLDMEADOW, J. A., SANTOS, I. M., TOWLER, J., BURT, D. M., AND YOUNG, A. W. Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition 127*, 1 (2013), 105–118.

[33] TODOROV, A., MANDISODZA, A. N., GOREN, A., AND HALL, C. C. Inferences of competence from faces predict election outcomes. *Science 308*, 5728 (2005), 1623–1626.

[34] TODOROV, A., OLIVOLA, C. Y., DOTSCH, R., AND MENDE-SIEDLECKI, P. Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology 66* (2015), 519–545.

[35] VAN DER MAATEN, L., AND HINTON, G. Viualizing data using t-sne. *Journal of Machine Learning Research 9* (11 2008), 2579–2605.

[36] WILSON, J. P., AND RULE, N. O. Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science 26*, 8 (2015), 1325–1331.

[37] ZEBROWITZ, L. A., KIKUCHI, M., AND FELLOUS, J.-M. Facial resemblance to emotions: group differences, impression effects, and race stereotypes. *Journal of Personality and Social Psychology 98*, 2 (2010), 175.

[38] ZHONG, Y., SULLIVAN, J., AND LI, H. Face attribute prediction using off-the-shelf cnn features. In *2016 International Conference on Biometrics (ICB)* (2016), IEEE, pp. 1–7.

[39] ZWILLINGER, D., AND KOKOSKA, S. *Standard Probability and Statistics Tables and Formulae*. Chapman & Hall/CRC, 2000.