NEURAL NETWORK COMPUTABILITY OF FACE-BASED ATTRACTIVENESS

Joshua Chauvin, Marcello Guarini
Department of Philosophy, University of Windsor, 401 Sunset, Windsor, ON, Canada

Christopher Abeare
Department of Psychology,University of Windsor, 401 Sunset, Windsor, ON, Canada

Keywords: Confidence, Face-based attractiveness, Face-based personality assessment, Face-based sex classification, Prototypicality effects, Intraclass correlation (ICC).

Abstract: In this work we have explored facial attractiveness as well as sex classification through the application of feed-forward artificial neural network (ANN) models. Data was collected from participants to compile a face database that was later rated by human raters. The neural network analyzed facial images as pixel-data that was converted into vectors. Prediction was carried out by first training the neural network on a number of images (along with their respective attractiveness ratings) and then testing it on new stimuli in order to make generalizations. There was strong intraclass correlation (ICC) and agreement between the neural network outputs and the human raters on facial attractiveness. This project's success provides novel evidence for the hypothesis that there are objective regularities in facial attractiveness. In addition, there is some indication that the confidence with which sex classification is performed is related to attractiveness. This paper corroborates the work of others that suggests facial attractiveness judgments can be learned by machines.

1 INTRODUCTION

To what extent can an artificial neural network (ANN) be trained to mimic human performance on facial attractiveness classification? Can ANNs learn to make human-like personality judgments? Could an ANN, trained to do sex classification, provide any evidence in support of the view that averageness (or prototypicality) is a contributing factor to attractiveness? This paper presents the preliminary results of a research project that engages the preceding research questions.

While facial attractiveness is recognized almost instantaneously (Locher et al, 1993), and personality characteristics are said to be assessed within a tenth of a second of seeing an unfamiliar face (Highfield, 2009), researchers are only beginning to explore neural network modeling of these human evaluations. The notion that beauty, namely facial attractiveness, is simply "in the eye of the beholder" has been effectively challenged and confronted with a "data-driven" (Eisenthal et al, 2006), or rather a biologically inspired, explanation for beauty.

Despite historic and cross-cultural differences in overall conceptions of beauty, assessments of facial attractiveness have been, on the whole, consistent throughout the world (Cunningham et al, 1995). Attributes such as facial averageness (Langlois & Roggman, 1990; Rhodes et al, 1999), facial symmetry (Grammer & Thornhill, 1994; Rhodes et al, 1999), sexual dimorphism and facial feminization (Perrett et al, 1998) are just some of the important features thought to aid in determining whether or not a face is considered attractive. Furthermore, evidence indicates that people not only judge an attractive individual to have more positive personality characteristics than an unattractive one (DeSantis & Kayson, 1997), they also tend to feel more personal regard and ascribe more power and competence to individuals they find physically attractive (Feingold, 1992; Fiske, 2001). For example, university professors are less likely to be blamed when a student receives a poor grade, and

Chauvin J., Guarini M. and Abeare C. (2009).
NEURAL NETWORK COMPUTABILITY OF FACE-BASED ATTRACTIVENESS.
In *Proceedings of the International Joint Conference on Computational Intelligence*, pages 473-479
DOI: 10.5220/0002322504730479

are more likely to be rated as better teachers if they are judged by the students to be more attractive (Romano & Bordieri, 1989).

Since there appears to be congruency among cultural representations of facial attractiveness, there is a strong likelihood that there may also be some biological criteria that guide such judgments. Given the preceding, it would seem plausible that a neural network, serving as a very powerful pattern classifier, could learn to recognize what humans find attractive, and effectively reproduce and generalize these assessments.

In previous attempts to model attractiveness, manually derived measurements between features as inputs were used and found to be successful. In contrast to this, researchers have extracted image factors associated with facial attractiveness from ratings of those images, and then designed a neural network to train and generalize based on those factors with strong correlations to human raters (Bronstad et al, 2008). Averaging, morphing digital images, and geometric modeling have been used in other work to construct attractive faces. Like Eisenthal et al (2006) and Bronstad et al (2008), we have not attempted to morph or construct attractive faces. Instead, we have used largely unmodified faces in order to retain nearly all aspects of face-based attractiveness assessments. Pixel-based images were inputted into an ANN –an approach that has been largely successful for other types of facial judgments, such as emotion classification (Dailey et al, 2002), sex classification (Cheng et al, 2001), and race categorization (Furl et al, 2002).

Using the images themselves, we try to train and test an ANN on attractiveness ratings as determined by human raters. We also train a network to carry out sex classification in order to determine if confidence in male and female images plays a role in attractiveness ratings. Initial results on training an ANN on personality features will not be discussed herein since they were based on raw data that is yet to be analyzed fully. Further analysis of that data will be reserved for discussion in a future paper.

2 DATA COLLECTION METHODS

2.1 Participants

There were two separate groups of participants investigated during data collection. For the first group, image data was collected on 100 undergraduate students (54 females and 46 males), aged 18 – 30 (mean = 22 and mode = 20 years), along with personality data for assessment. A second group of 104 undergraduate students (52 females, 47 males, 1 self-classified as "other" and 4 with missing data) aged 18-61 (mean = 23 and mode = 20 years) rated the image data collected for attractiveness and personality traits. Both samples were noticeably diverse, with a mix of racial and ethnic backgrounds. Participants were recruited on a volunteer basis through the university psychology participant pool during separate semesters and were not allowed to participate in both parts of the study (i.e., the 'image collection stage' and the 'image rating stage' were exclusive). All participants provided informed consent, and course credit was given for participating in the study.

2.2 Description of Measures

2.2.1 Procedure for Image Collection

Participants who volunteered for the first part of the study were asked to fill out a consent form specific to having their picture taken. After consenting, participants were asked to fill out a brief demographics form. Participants were then photographed and asked to complete a shortened version of the Big Five Inventory (BFI) personality test (John & Srivastava, 1999).

2.2.2 Image Ratings

Those who participated in the second part of the study were asked to fill out a brief demographics form and to take part in a short personality questionnaire (the BFI) after having consented.

Subsequently, a questionnaire with the images collected from the prior phase was presented in DirectRT, a computerized stimulus presentation program, and the participants were instructed to evaluate the images according to ten propositions that coincide with the dimensions of personality measured in the BFI. (The BFI measures the "Big 5" personality traits, which include: Extraversion, Agreeableness, Conscientiousness, Neuroticism and Openness.) Additionally, participants were asked to assess the facial attractiveness of the presented image. Before evaluating the individual faces on the various dimensions, participants were prompted to indicate whether or not they recognized the individual they were rating. All images that were recognized by the participants were not evaluated, and participants were required to move on to the next image in an attempt to ensure zero acquaintance.

Attractiveness ratings were evaluated using a ten-level Likert scale (i.e. 1= Very Unattractive, 3= Unattractive, 5 = Somewhat Unattractive, 6 = Somewhat Attractive, 8 = Attractive, and 10 = Very Attractive), while the other ten personality questions were formatted according to a typical five-level Likert scale (i.e .1= Strongly Disagree, 2= Disagree a little, 3= Neither Agree Nor Disagree, 4= Agree a little, and 5= Strongly Agree). All questions were asked in a randomized order aside from attractiveness, which always appeared at the end of the list as the eleventh item. The order of test administration was counterbalanced and randomized with the purpose of controlling for order effects. All of the above mentioned methods were approved by the University of Windsor Research Ethics Board.

2.3 Images

A total of 100 photographs were taken (54 women and 46 men), yielding 99 usable images. One image was removed from the dataset due to image file corruption. Lighting and background were held constant, and a 3.1 mega pixel camera was set in the same position for every participant. Each image was converted to 8 bit grey scale (i.e. 256 shades of grey) and reduced to 180 x 256 pixels. These grey scale images were the ones reviewed by the raters.

Given that in real life attractiveness assessments are made under less than perfect conditions, accessories such as glasses, headbands, hair clips and headscarves were allowed to remain on in order to assess whether accurate neural network attractiveness predictions would still be possible.

The images presented to the neural networks remained as 256 shades of grey. However, to minimize training times and maximize the number of training runs, the networks were presented with 64 x 91 pixel images (the reduction preserved the aspect ratio).

3 NEURAL NETWORKS

3.1 Architectures

PDP++ 3.1 was used to create, train, and test all ANN simulations. Fully interconnected feed-forward networks were used in all work discussed in this paper. The generalized delta rule was used for all training. Images were converted into vectors suitable for input to the ANNs. In all cases, the networks had 5824 input units, one for each pixel of the image (each image was 64 x 91 pixels). The value of each unit varied from 0 to 255, corresponding to the 256 shades of grey in the images (see Figure 1).

The number of hidden units in the ANNs varied with the tasks they were asked to perform. We found that with respect to rating attractiveness, networks with 60 hidden units performed best. With respect to the task of classifying images into either male or female, networks with 120 units worked best.

All networks discussed herein contained 1 output unit.

For attractiveness rating networks both nonlinear sigmoid and radial basis activation functions were used. All training and testing results discussed in this paper refer to sigmoid networks since results for attractiveness rating using the radial basis function were inferior to networks using the sigmoid activation function.

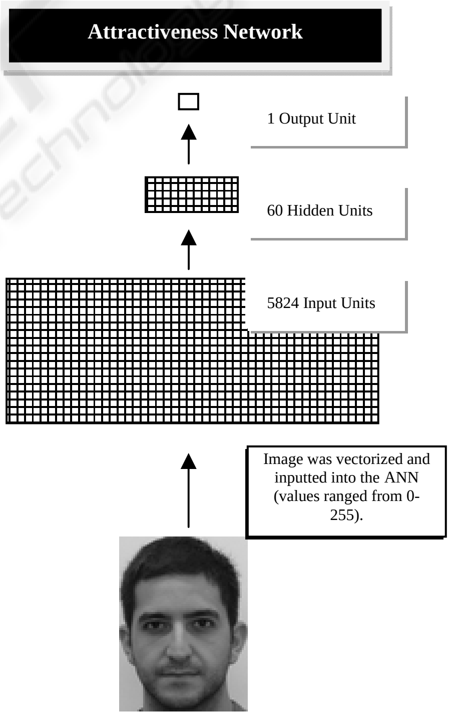

Attractiveness Network
1 Output Unit
60 Hidden Units
5824 Input Units
Image was vectorized and inputted into the ANN (values ranged from 0-255).

Figure 1: Attractiveness Network. Visual depiction of fully interconnected feed-forward neural network model (not to scale). Image is 64 x 91 pixels and is taken from the sample of participants.

3.2 Training

3.2.1 Training the Attractiveness Network

For training a network to make predictions about facial attractiveness, the desired output for an image that scored 3 out of 10 was set to 0.3. The desired output for an image that rated 4 out of ten was 0.4, and so on to images that scored 8 out of ten, where the desired output was set to 0.8. (Since none of the images averaged scores of 1, 2, 9, or 10, desired output values of 0.1, 0.2, 0.9, and 1.0 were never used.)

SUM training and COUNT training were used. Since there is only one output neuron for the net, the sum of squared error (sse) for the output layer is simply the squared error (se) of the output neuron. In SUM training we set the sse for the entire training batch (i.e., the error level at which to terminate training) to a number of different values, finding that values around 0.35 worked best.

In COUNT training, we set the desired se at 0.0025 (or less) for each image, and set the simulator to count the number of images having that level of error, terminating training when 0 images had errors. With these specifications we could not get the network to train. When we tolerated more error, terminating training with 3 errors, the network trained, but it did not generalize as well as networks trained using SUM training. Using the COUNT method, we experimented with tolerating varying levels of error per image and varying levels of error tolerance for the training set, but we never achieved the same level of success as we did with SUM training.

We discovered that with both SUM and COUNT training there were four images in the training set that consistently failed to train over hundreds of runs. We removed these images from the original training set of 66, yielding a training set for attractiveness of 62 images and a testing set of 33 images. Even using SUM training on 62 images there are errors, but the errors vary from training on one set of initial weights to other randomly selected sets of initial weights. Results discussed below with respect to predicting attractiveness refer to training with 62 cases and testing/predicting with 33.

3.2.2 Training the Sex Classification Network

For training purposes, the desired output for all female images was set to 0; the desired output for all male images was set to 1. Again, we used both SUM and COUNT methods. When using the COUNT method, we were able to train the network to successfully classify all 99 images. This was done by setting the error target for each image to less than 0.25. The simulator was set to count the number of images for which the network had errors and to terminate training when it had 0 errors. (Any male image with an output of above 0.5 was considered successfully classified, and any image of a female below 0.5 was considered successfully classified.) 120 hidden units were required to achieve a network that trained on *all* 99 cases. Networks with fewer hidden units consistently failed to train.

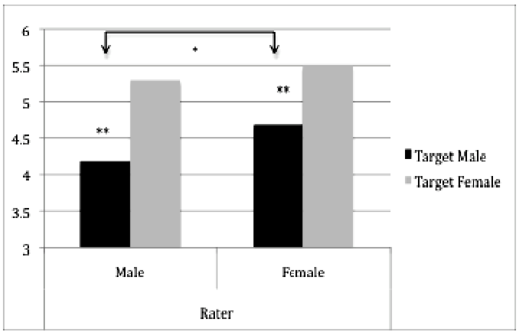
When using the SUM method, we set training to terminate when the sse for the entire batch of 99 images was less than 2.5. While the network did train to that level of error tolerance for the whole batch, there were still errors with individual images. To get the level of success we did manage to achieve, again, 120 hidden units were required. We experimented with different levels of error tolerance without improving results. When the sse for the entire batch was set below 2, we could not get the network to train.

4 RESULTS

4.1 Participant Ratings

Mean attractiveness ratings for each face ranged from 2.27 to 7.83 with a mean of 4.97 (SD = 1.11). Missing values for facial attractiveness ratings were replaced with the mean for that target face. Attractiveness ratings were calculated by sex of rater and sex of target (See Figure 2). There was a moderate correlation between the ratings of female and male faces, r = .59. Males and females rated females as most attractive. Average male ratings of females (mean = 5.29 SD = 1.02) was higher than male ratings of males (mean = 4.19, SD = 1.34), $t_{(44)} = 5.21$, $p < .001$. Average female ratings of females (mean = 5.50, SD = 1.09) was also higher than for males (mean = 4.69, SD = 1.08), $t_{(50)} = 11.60$, $p < .001$, however, males were rated higher by females than by males, $F_{(1,98)} = 4.07$, $p < .05$.

Reliability was assessed through intraclass correlation (ICC) as an index of absolute rater agreement (Shrout & Fleiss, 1979). The two-way random effects ICC for the sample ($ICC_{(2,100)} = .962$) reflected a high level of absolute inter-rater agreement. In order to be consistent with reporting practices of previous studies, internal consistency reliability was calculated, Cronbach's α = .978. Separate ICCs were calculated for males ($ICC_{(2,48)}$ = .950) and females ($ICC_{(2,48)}$ =.969) and were comparable to each other and to the overall ICC.



* = p < .05
** = p < .001

Figure 2: Mean Attractiveness Ratings by Sex of Rater and Sex of Target.

4.2 ANN & Attractiveness Ratings

After training on attractiveness ratings for 62 images, the network's performance was assessed by testing on 33 novel cases. There was a substantial degree of agreement between the neural network output on novel cases and the participant ratings. The average ICC for the four simulations was $ICC_{(2,32)}$ = .696, demonstrating that the scores produced by the neural network were closely related to the scores produced by the participant raters (See Table 1 for values for all four simulations). More specifically, 56% of the neural networks ratings were an exact match with the participant ratings and an additional 29% were within one point of the participant ratings making for 85% of the neural network's ratings falling within one point of the participant ratings.

Table 1: Pearson's Correlation Coefficients and Intraclass Correlation Coefficients (ICC) between Raters' and Neural Network Simulations' Attractiveness Ratings.

Simulation	Pearson's Correlation	ICC
1	.608	.677
2	.612	.707
3	.612	.707
4	.559	.693
Mean	0.598	0.696

4.3 ANN & Sex Classification

As indicated above, COUNT training was used to achieve 100% success in classifying all 99 images as either male or female. The closer the output for a male image was to 1, the lower its se. The closer the output for a female image was to 0, the lower its se. The closer the output for an image is to 0.5 (for either male or female), the greater its se. We took images with a lower se to be more confidently classified as male or female (with respect to the set of 99 images) since higher se means the image is approaching the opposite classification. After training a network using COUNT to correctly class all males above 0.5 and all females below 0.5, we compared the se of the images in the sex classification task with the attractiveness ratings of the images. If attractiveness increases as confidence increases, and a decrease in se in the sex classification task means an increase in confidence, then one would expect that as se in the sex classification task decreases, attractiveness increases. What follows is some of the evidence we found for this trend.

In one training run of the sex classification net, we received a very impressive result. We used the sex classification se for each image (processed by a fully trained network) to compute the mean sex classification se for images rating 8/10; we did the same for images rating 7/10, and so on down to 3/10. It turned out that the lowest mean se (or highest mean confidence) in sex classification was for images scoring 8/10. The second lowest mean se (or next highest mean confidence) was for images scoring 7/10; and the pattern continued right down to 3/10. While very impressive, the finding at that level of detail was not robust. We did an additional four training runs (starting with randomly selected weights every time) and did not achieve the same results (e.g., sometimes 7/10s had lower se than 8/10s). However, we did find a result consistent over all five training runs. If we take the mean sex classification se of all images with ratings of 3/10, 4/10, and 5/10 (the low end) and compare them with the mean sex classification se for all images with ratings of 6/10, 7/10, and 8/10 (the high end), it turns out that the mean se for the low end is higher than the mean se for the high end in *all* five training runs. In other words, on average, the ANN more confidently assigned male or female classifications to images that scored in the high end of attractiveness than to those that scored in the low end.

5 DISCUSSION

We have presented a neural network model, trained on a diverse sample of images of both males and females (with their respective human ratings), to predict facial attractiveness means with a high degree of correlation and agreement with human raters. This study helps to reinforce the claim that attractiveness assessments are data-driven, and further expands on existing research using computational modeling to make facial attractiveness judgments. Given a larger dataset, it may be possible to create a neural network that is capable of producing human-like evaluations with stronger correlations and agreement with human raters.

In addition to learning facial attractiveness, we have trained an ANN to distinguish between both males and females, and found some evidence that would suggest confidence plays a role in sex classification. *If* it turns out that these confidence ratings correspond to prototypicality or averageness – the more confident the network is that an image is male (or female) the more prototypically male (or female) it is – then we would have an especially interesting result. This speaks to a larger and more difficult question we have insufficient room to explore at this point: *why* is prototypicality a contributing factor to facial attractiveness? If an ANN, in solving for sex classification, yields prototypical male and female outputs in a way that at least roughly corresponds to attractiveness ratings, then one starts to wonder about the following hypothesis: *the contribution of prototypicality to facial attractiveness could be a neurocomputational consequence of mastering the task of male-female facial classification.* In other words, the contribution of prototypicality to attractiveness may "fall out of" the solution to male-female classification (of course, as literature surveyed in the introduction suggests, prototypicality is only one of several contributors to attractiveness). That said, any link between confidence ratings and prototypicality needs to be independently motivated. Moreover, much more work is required in neuropsychology and computational modeling to examine the preceding hypothesis, but it is at least worth mentioning at this point.

All training of the sex classification network made use of the sigmoid activation function. We have not yet trained sex classification networks using the radial basis function. This is an important consideration for future work since the radial basis function may make it easier than the sigmoid function to motivate a link between confidence ratings and prototypicality.

In conclusion, this work has produced useful results. There were significant correlations with human ratings of attractiveness despite the ostensible difficulty of this computational task. Corroborating other research, it would seem that there are grounds to believe that human assessments of facial attractiveness can be learned by a machine.

ACKNOWLEDGEMENTS

The authors would like to thank Kaitlin Tremblay and Emrah Eren for helping to compile and compute some of the data. An additional thanks is extended to the participant in the above picture for allowing the use of his image as well as to Kate Hargreaves for editing.

REFERENCES

Bronstad, P. M., Langlois, J. H., & Russell, R. (2008). Computational models of facial attractiveness judgments. *Perception, 37*(1), 126.

Cheng, Y. D., O'Toole, A. J., & Abdi, H. (2001). Classifying adults' and children's faces by sex: Computational investigations of subcategorical feature encoding. *Cognitive Science: A Multidisciplinary Journal, 25*(5), 819-838.

Cunningham, M. R., Roberts, A. R., Barbee, A. P., Druen, P. B., & Wu, C. H. (1995). 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), 261-279.

Dailey, M. N., Cottrell, G. W., Padgett, C., & Adolphs, R. (2002). EMPATH: A neural network that categorizes facial expressions. *Journal of Cognitive Neuroscience, 14*(8), 1158-1173.

DeSantis, A., & Kayson, W. A. (1997). Defendants' characteristics of attractiveness, race, and sex and sentencing decisions. *Psychological Reports, 81*(2), 679-683.

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

Feingold, A. (1992). Gender differences in mate selection preferences: A test of the parental investment model. *Psychological Bulletin, 112*(1), 125-139.

Fiske, S. T. (2001). Effects of power on bias: Power explains and maintains individual, group, and societal disparities. *The use and abuse of power: Multiple perspectives on the causes of corruption,* 181-193.

Furl, N., Phillips, P. J., & O'Toole, A. J. (2002). Face recognition algorithms and the other-race effect:

Computational mechanisms for a developmental contact hypothesis. *Cognitive Science: A Multidisciplinary Journal, 26*(6), 797-815.

Grammer, K., & Thornhill, R. (1994). Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *Journal of Comparative Psychology, 108*(3), 233-242.

Highfield, R., Wiseman, R., & Jenkins, R. (2009). In your face. *New Scientist, 201*(2695), 28-32.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research, 2*, 102-138.

Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science, 1*(2), 115-121.

Locher, P., Unger, R., Sociedade, P., & Wahl, J. (1993). At first glance: Accessibility of the physical attractiveness stereotype. *Sex Roles, 28*(11), 729-743.

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., et al. (2002). Effects of sexual dimorphism on facial attractiveness. *Foundations in Social Neuroscience*, 937.

Rhodes, G., Sumich, A., & Byatt, G. (1999). Are average facial configurations attractive only because of their symmetry? *Psychological Science, 10*(1), 52-58.

Romano, S. T., & Bordieri, J. E. (1989). Physical attractiveness stereotypes and students’ perceptions of college professors. *Psychological Reports, 64*(3 Pt 2), 1099–1102.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull, 86*(2), 420-428.