Genuine Personality Recognition from Highly Constrained Face Images

Fabio Anselmi[1], Nicoletta Noceti[2(✉)], Lorenzo Rosasco[1,2], and Robert Ward[3]

[1] LCSL, Istituto Italiano di Tecnologia and Massachusetts Institute of Technology, Cambridge, USA
{fabio.anselmi,lorenzo.rosasco}@mit.edu
[2] Università degli Studi di Genova, Genoa, Italy
nicoletta.noceti@unige.it
[3] Bangor University, Bangor, UK
r.ward@bangor.ac.uk

Abstract. People are able to accurately estimate personality traits, merely on the basis of "passport"-style neutral faces and, thus, cues must exist that allow for such estimation. However, up to date, there has been little progress in identifying the form and location of these cues.

In this paper we address the problem of inferring true personality traits in highly constrained images using state of art machine learning techniques, in particular, deep networks and class activation maps analysis.

The novelty of our work consists in that, differently from the vast majority of the current and past approaches (that refer to the problem of consensus personality rating prediction) we predict the genuine personality based on highly constrained images: the target's are self ratings on a validated personality inventory and we restrict to passport-like photos, in which so-called controllable cues are minimized.

Our results show that self-reported personality traits can be accurately evaluated from facial features. A preliminar analysis on the features activation maps shows promising results for a deeper understanding on relevant facial cues for traits estimation.

Keywords: Five-Factor Model · Convolutional neural networks · Class activation maps

1 Introduction

Psychological studies of "thin slices" of behaviour investigate the accuracy of social judgements made by observers on the basis of minimal information, usually nonverbal and unintentional cues emitted by the target person being judged. For example, ratings of university lecturers, made on a basis of silent 2-s video clips, significantly correlated with the lecturers' end-of-semester student evaluations [1]. Of particular interest, observers are able to accurately estimate personality traits on the thinnest of slices, merely on the basis of "passport"-style neutral

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11751, pp. 421–431, 2019.
https://doi.org/10.1007/978-3-030-30642-7_38

face images [13]. "Accuracy" here refers to the agreement between self-ratings of personality, as made by the targets themselves, and the ratings of observers.

The long-term goal of our work is to gain a deeper understanding on the cues used by observers to make accurate judgements of personality from these highly constrained face images. To date, there has been little progress in identifying the form and location of these cues [21,22]. However, given that observers can make accurate judgements of true personality, these cues must exist, and recent computational advances offer promise. From a computational perspective, before understanding the *what* and *where* of these cues, we need to first assess the feasibility of the underlying task, that is, estimating true personality of people from a single highly constrained image of their face. That is our task here, where, given a set of face images annotated with personality characteristics, we employed a pre-trained model obtained with a well-known convolutional neural network, the VGG16 [16], to extract the relevant features from each image. Then, we trained a fully connected network for personality characteristics regression.

In the experiments, we considered a dataset of still images annotated with the personality scores of the observed subject. Targets were explicitly required to adopt neutral expressions, hair back, cosmetics, glasses and jewelry removed (according to e.g. [13]). The targets' genuine personalities were measured through self-rating with validated five-factor personality inventories, i.e. NEO-IPIP or mini-IPIP (see e.g. [5]).

Our work differs from previous approaches in several ways. First, we are predicting the *true* (or actual, genuine) personality of target persons. The actual personality is defined as the target's self ratings on a validated personality inventory. In contrast, the vast majority of approaches in the computing science fields refer to the problem of *apparent* (or consensus) personality rating prediction [3,7,10,12,19], as made by external observers. While the consensus rating is relatively easy to collect [2], actual personality ratings are generally more laborious and expensive to obtain.

A second aspect of diversity is that, unlike other work (see e.g. [6,18]) we are using highly constrained images, similar to passport photos, in which so-called controllable cues are minimized. Controllable cues include all that aspects of the face image which can be readily modified by the target to create different personality impressions, influencing the consensus personality rating. In this respect, our work shares similarities with [11] who looked for correlation between 3D face structure and personality. However, our stimuli are notably different in that they are 2D and include both shape and surface information.

In summary, to the best of our knowledge, this paper represents the first attempt to address inference of true personality in highly constrained images.

The remainder of the paper is organized as follow. In Sect. 2 we describe our dataset, while in Sect. 3 details on our approach for the personality scores prediction (Sect. 3.1) and on the experimental assessment (Sect. 3.2) are reported, with a preliminar analysis on class activation maps (Sect. 3.3). Sect. 4 is left to a final discussion.

Fig. 1. Average images of a few samples of female (left) and male (right) faces in the dataset.

2 The Data

In this work we employ a dataset acquired in-house consisting of 997 still images depicting the face of a target individual. People have been asked to avoid such cues one can voluntary control – as hairstyle, jewellery, facial expressions, and cosmetics – that in general may affect personality judgements. In Fig. 1 we report two average images of a few samples of female (left) and male (right) faces in the dataset[1] which includes in total 604 female and 393 male faces.

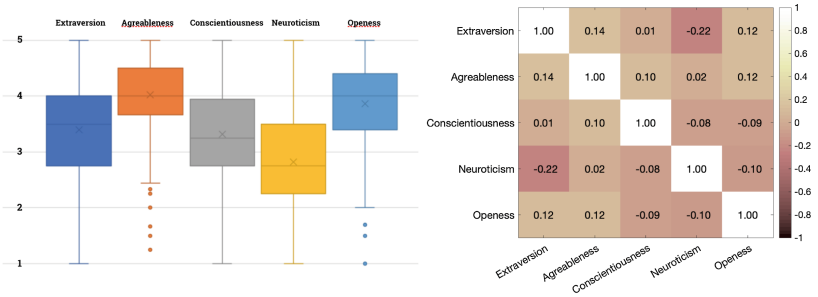
For each target a standard five-factor model [FFM] personality inventory – NEO-IPIP or mini-IPIP – has been proposed, to finally collect a six dimensional vector containing the score for each personality trait plus the information on the gender. To the authors knowledge it is the first time that a dataset with such characteristics is analyzed.

The FFM is defined in terms of five different dimensions (see [9]):

- *Extraversion* vs. *Introversion*, scoring from sociable, assertive, playful to aloof, reserved, shy.
- *Agreeableness* vs. *Disagreeable*, scoring from friendly, cooperative to antagonistic, fault-finding.
- Conscientiousness vs. *Unconscientious* scoring from self-disciplined, organised, to inefficient, careless.
- *Neuroticism* vs. *Emotional stability*, scoring from calm, unemotional to insecure, anxious.
- *Openess* vs *Closed to experience*, scoring from intellectual, insightful to shallow, unimaginative.

Before proceeding with the presentation of our methodology, it is worth discussing some of the properties of the dataset with a brief statistical analysis. A first aspect to be mentioned is the fact that although in principle the range of values that the scores can assume is (1–5) (with maximal resolution of 0.25), the actual distributions of the estimated scores are uneven (see Fig. 2(a)).

[1] For privacy issues we can not show the original images.



(a) (b)

Fig. 2. Left: distribution of the FFM traits scores in the dataset. Right: correlation coefficients between the FFM traits.

A second aspect to be considered refers the presence of correlations among the different traits. We evaluated the correlation with the Spearman's Rank Coefficient and report in Fig. 2(b) the results we obtained between the scores of the whole dataset. According to the classification in [8] very weak correlation can be noticed overall among traits, with only one weak correlation between *Extraversion* and *Neuroticism* ($\rho = -0.22, P < 10^{-11}$). Notice however that the null hypothesis (i.e. there is no correlation between a specific pair of traits) can not be rejected at the significance level of 0.05 for the pairs *Extraversion-Conscientiousness* ($\rho = 0.02, P = 0.54$) and *Agreeableness-Neuroticism* ($\rho = 0.02, P = 0.53$).

As it is well known that personality traits may significantly differ in male and female populations, we verified this aspects on the dataset. Performing a statistical comparison between female and male samples using a Two-Sample TTest we assessed that the only trait presenting no significant difference in a statistical sense is *Extraversion* ($P = 0.92$), while for all the others the test reveals a strong separation between the two sample sets (in all cases $P < 10^{-4}$). A visualization of the approximated trait distributions, represented with histograms, is also reported in Fig. 3.

3 Estimating the Five-Factor Model from Images

In this section we discuss the methodology we applied to estimate the scores describing the Five-Factor Model of an individual from a single highly-constrained image of his/her face. To this purpose, we casted the estimation problem to a regression task where we want to learn the mapping between an appropriate representation of the face image and the annotated scores associated with it.

The images in the dataset presented in the previous section have been processed with a face detector [20], to precisely identify the image portion corresponding to the face, and then converted to grayscale. The resulting segmented

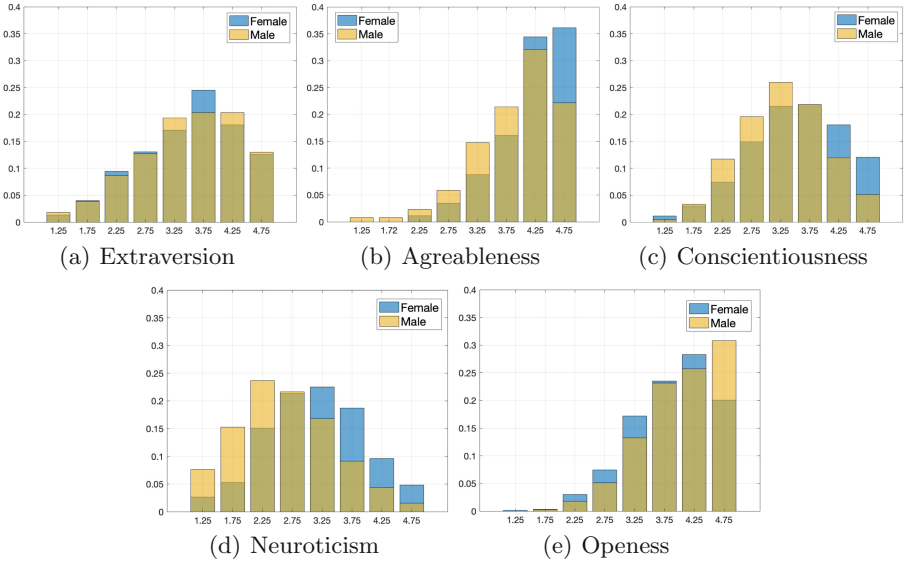


Fig. 3. Histograms describing the distribution of the 5 personality traits in the female and male samples.

images have finally been resized to a fixed size of 224 × 224 pixels. Then we extracted the relevant image features using the convolutional part of a pre-trained Neural Network [17]. Finally we used them as input to a fully connected layer to predict the Five-Factor scores for each target individual.

All the computational models are implemented using Keras and Tensor Flow [4].

3.1 Personality Traits Regression with Pretrained VGG16 Network

To extract the relevant features from each image we used the convolutional part of a pre-trained VGG16 convolutional neural network, [17] – pre-trained on the Imagenet dataset [15].

The output of this first part of the architecture is a set of feature vectors of 25088 components that we used to train a fully connected network composed of three dense layers, two of 100 units and a last one with size that depends on the specific task we solve. Indeed, we explored two main regression tasks. With the first, we trained the network to learn the mapping to each trait independently from the others. We refer to this task as "single personality trait regression", and this corresponds to using a last layer composed of a single unit (size $1 \times 1 \times 1$). The second task considers instead the possibility of exploiting possible hidden correlations among traits, by learning them as a whole with a vectorial regression model. We will refer to it as "full personality traits regression", for which the last layer is composed of 6 units (size $1 \times 1 \times 6$). A visual sketch of the deep architecture we finally adopted is reported in Fig. 4.

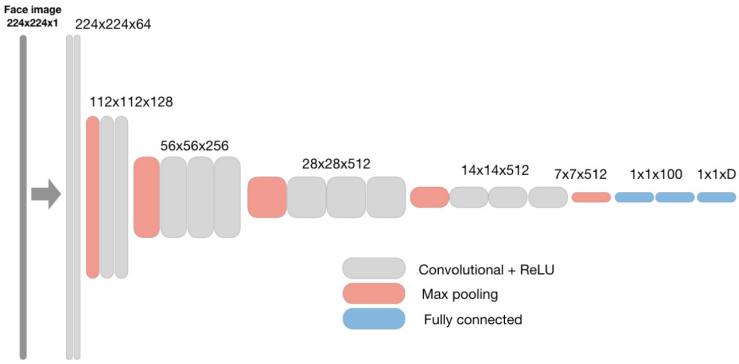


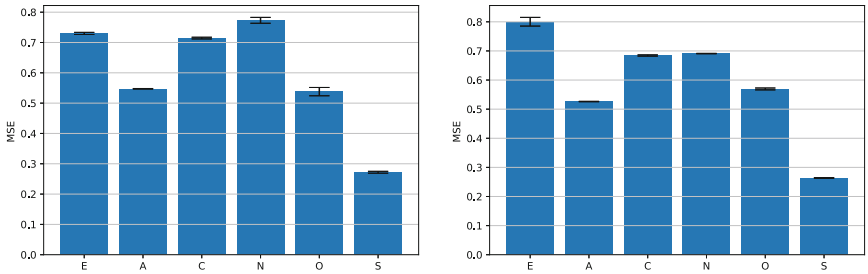
Fig. 4. A visualization of the architecture we adopted in our experiments. The convolutional part of the VGG-16 deep network is followed by dense layers, the latter having different size depending on the specific task to be addressed: D = 1 for single personality trait regression, while D = 6 for full personality traits regression (see text for details).

The training has been performed on the 80% of the data samples, while the remaining 20% has been employed for evaluation only. At each of the 100 epochs, after a random permutation of the images, the training set was further split into 70% and 30% validation set. To avoid overfitting we used a dropout regularization within the dense layers with rate 0.3. An Adam optimizer was chosen with a batch size of 400 and a learning rate of 0.001 to minimize the mean square error (MSE) loss. The full protocol was replicated 10 times to have a statistics on the MSE on the test set.

3.2 Experimental Assessment

In this section we discuss the results of the experimental analysis we carried out. In Fig. 5 we report the average Mean Square Error (MSE) obtained on the test set for 10 random data splitting, using the architecture depicted in Fig. 4. More specifically, we compare the effects of learning each trait independently from the others (Fig. 5(a)) with the use of a vectorial regression to learn all the traits as a whole (Fig. 5(b)). A first observation refers to the fact that overall the MSEs we obtained are very promising for all the traits (as a reference consider the reported MSE on sex (S) an easy individual characteristic to predict). In this evaluation we implicitly consider the fact that the annotation we use as a reference is result of a quantization from outcomes of self-reported questionnaires, influenced by the individual subjectivity and thus prone to error. The results we obtained are in line with performance reported in works sharing some contacts points with ours although grounding on different motivations (as e.g. [23]).

A visual comparison with the results obtained for the full personality traits regression task highlights uneven effects, in the sense not all the traits seem to benefit from the full vector regression. Table 1 reports in the second and third



(a) Single personality traits regression and sex (S).

(b) Full vector personality traits regression and sex (S).

Fig. 5. Average mean square error on the test set using the architecture in Fig. 4 with D = 1 (left) and D = 6 (right). Error bars refer to $N = 10$ repetitions.

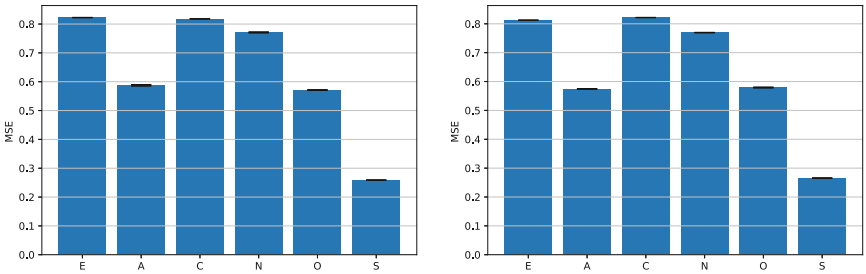
colums the MSE values for, respectively, single and full traits regression. For each trait, we highlighted in bold the best MSE among the two tasks.

Comparable results have been obtained with a recent kernel-based approach for large-scale datasets [14], an alternative method we used to asses the consistency of our results. The method builds on the use of Kernel Ridge Regression with a gaussian kernel on the same set of features vectors. Figure 6 reports a visual impression of the MSE obtained with the method on the same regression tasks considered above, and the average values are also reported in Table 1 (forth and fifth columns) for a comparison. The values suggest non significant statistical correlation is present among traits. It is in particular interesting to note that the gender of the target individual seems to play a minor role in the score estimate, although statistical evidences that the traits are different in female and male population have been assessed.

The overall results support our methodology and show how self-reported personality traits can be evaluated from single, strongly constrained face images. This result suggests that the features at the basis of the representation may in fact helps in understanding the cues that reveals personality traits. In the next section we thus discuss a preliminary analysis in this direction.

3.3 Activation Maps: A Preliminary Analysis

The assessment of the regression task that proves how it is possible to estimate the Five-Factor model from one single image of an individual, allows us to further investigate on the cues of the face enabling this ability. Since we use highly constrained face images, the effects of controllable cues that might influence the target personality judgment from an observer (and it is fair to assume the same for a computational model) can be neglected. We can thus assume that the visual features showing strong responses for the Five-Factors model estimation are in fact face cues inherently providing essential judgement information.



(a) Single personality traits regression and sex (S).

(b) Full vector personality traits regression and sex (S).

Fig. 6. Average mean square error on the test set using the Falkon method [14] (error bars refer to $N = 10$ repetitions).

Table 1. Average mean square errors and standard deviations obtained on the test set for the single and full traits (plus sex (S)) regression models using different approaches (see text for details).

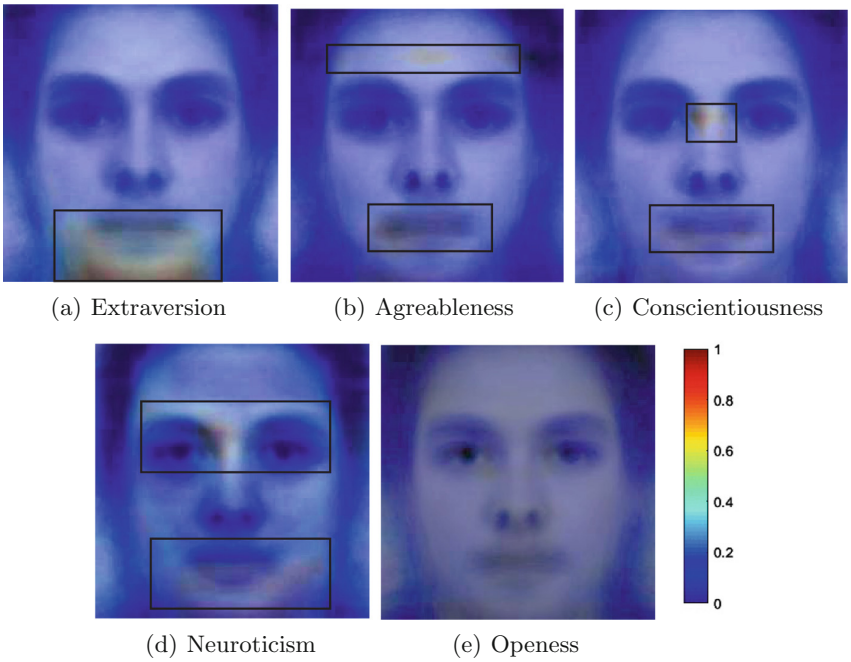
Big5	VGG16 single	VGG16 full	Falkon single	Falkon full
E	**0.7300** ± 0.0035	0.8002 ± 0.0147	0.8377 ± 0.0011	**0.8243** ± 0.0014
A	0.5473 ± 0.009	**0.5261** ± 0.0006	0.5723 ± 0.0010	**0.5654** ± 0.0003
C	0.7145 ± 0.0028	**0.6842** ± 0.0021	**0.8086** ± 0.0002	0.8119 ± 0.0003
N	**0.7731** ± 0.0097	0.9609 ± 0.0006	**0.7529** ± 0.0013	0.7675 ± 0.0004
O	**0.5379** ± 0.0137	0.5690 ± 0.0032	0.5745 ± 0.0006	**0.5733** ± 0.0017
S	0.2719 ± 0.0030	**0.2636** ± 0.0010	**0.2619** ± 6 $\cdot 10^{-5}$	0.2658 ± 9 $\cdot 10^{-5}$

Considering the well-assessed theory about deep features visualization in classification settings [16,24], to the purpose of feature visualization we converted our task to a multi-class classification problem. This choice allowed us to use off the shelf, well tested, algorithms like gradient weighted class activation maps (CAM) for classification (not regression). To this aim we quantized the traits score range into N (in our experiments empirically set to 6) intervals and trained a single fully connected classifier on the same feature vectors dataset as for the regression task. As a consequence, the architecture in Fig. 4 has been slightly modified replacing the very last fully connected layer with a soft-max layer.

To highlight the target image parts on which the network was focusing to make its prediction we used the gradient weighted class activation map (CAM, [16]).

Figure 7 shows the average heatmaps (overlaied to the original images) we obtained for the estimation of high values of the personality scores.

The highlighted regions of maximal signal in the figure show how well defined parts of the individual target faces are selected to score the individual traits. In the case of *Openness* the averaged signal was not significant. Of particular



(a) Extraversion (b) Agreableness (c) Conscientiousness

(d) Neuroticism (e) Openess

Fig. 7. Average of superposed heatmaps on each subject correctly predicted for the single trait class. Highlighted zones of maximal signal.

interest are the results for the *Neuroticism* trait, where the two most sexually dimorphic regions of the human face have been tagged: the jaw and the brow. Indeed, it is worth noticing that neuroticism is known to be sexually dimorphic (women higher than men, as also visible from the histograms in Fig. 7(d)), and deserves further study.

4 Discussion

Estimation of personal traits is important for designing personality-aware intelligent systems and the computational model beyond the estimation might allow to make a step towards the understanding and the characterization of the elements of faces to judge social traits. Considering that people are normally able to accurately estimate personality traits, merely on the basis of a "passport"-style neutral face, we hypothesized that cues must exist that allow for such estimation and we addressed the problem of inference of true personality.

To this aim we employed state of art machine learning techniques, in particular, deep convolutional networks and class activation maps analysis, to test our hypothesis and highlight specific face cues upon which personality traits can be inferred.

The novelty of our work consists in the fact that, differently from the vast majority of the current approaches that refer to the problem of apparent (or consensus) personality rating prediction, we predicted the genuine personality of target persons, as the target's self ratings on a validated personality inventory. Also, we focused on highly constrained images, in which so-called controllable cues are minimized.

Our results supported our methodology and hypothesis and show how self-reported personality traits can be accurately evaluated from the facial features. The class activation maps analysis further confirmed the feasibility of our approach showing, for example, that the two most sexually dimorphic regions of the human face, the jaw and the brow, have been correctly tagged by the network to infer the trait score. Our initial results are very promising and a more detailed analysis based on class activation maps will be the subject of further study.

Acknowledgement. This material is based upon work partially supported by the Italian Institute of Technology.

L. R. and F.A. acknowledges the financial support of the AFOSR projects FA9550-17-1-0390 and BAA-AFRL-AFOSR-2016-0007 (European Office of Aerospace Research and Development), and the EU H2020-MSCA-RISE project NoMADS - DLV-777826.

References

1. Ambady, N., Rosenthal, R.: Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. J. Pers. Soc. Psychol. **64**(3), 431 (1993)
2. Biel, J.I., Gatica-Perez, D.: The youtube lens: crowdsourced personality impressions and audiovisual analysis of vlogs. IEEE Trans. Multimed. **15**(1), 41–55 (2013)
3. Celli, F., Bruni, E., Lepri, B.: Automatic personality and interaction style recognition from Facebook profile pictures. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 1101–1104. ACM (2014)
4. Chollet, F., et al.: Keras (2015). https://github.com/fchollet/keras
5. Donnellan, M.B., Oswald, F.L., Baird, B.M., Lucas, R.E.: The mini-IPIP scales: tiny-yet-effective measures of the big five factors of personality. Psychol. Assess. **18**(2), 192 (2006)
6. Ferwerda, B., Schedl, M., Tkalcic, M.: Predicting personality traits with Instagram pictures. In: Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015, pp. 7–10. ACM (2015)
7. Ferwerda, B., Schedl, M., Tkalcic, M.: Using Instagram picture features to predict users' personality. In: Tian, Q., Sebe, N., Qi, G.-J., Huet, B., Hong, R., Liu, X. (eds.) MMM 2016. LNCS, vol. 9516, pp. 850–861. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-27671-7_71
8. Fowler, J., Cohen, L., Jarvis, P.: Practical Statistics for Field Biology. Wiley, Hoboken (2013)
9. Goldberg, L.R.: The structureof phenotypic personality traits. Am. Psychol. **48**, 26 (1993)
10. Guntuku, S.C., Qiu, L., Roy, S., Lin, W., Jakhetiya, V.: Do others perceive you as you want them to?: Modeling personality based on selfies. In: Proceedings of the 1st International Workshop on Affect and Sentiment in Multimedia, pp. 21–26. ACM (2015)

11. Hu, S., et al.: Signatures of personality on dense 3D facial images. Nat. Rep. **7**, 73 (2017)
12. Junior, J., et al.: First impressions: a survey on computer vision-based apparent personality trait analysis. arXiv preprint arXiv:1804.08046 (2018)
13. Kramer, R.S., Ward, R.: Internal facial features are signals of personality and health. Q. J. Exp. Psychol. **63**(11), 2273–2287 (2010)
14. Rudi, A., Carratino, L., Rosasco, L.: Falkon: an optimal large scale kernel method. In: Advances in Neural Information Processing Systems, vol. 30, pp. 3888–3898. Curran Associates, Inc. (2017)
15. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. **115**, 211–252 (2015)
16. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556
18. Sutherland, C.A., et al.: Personality judgments from everyday images of faces. Front. Psychol. **6**, 1616 (2015)
19. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. IEEE Trans. Affect. Comput. **5**(3), 273–291 (2014)
20. Viola, P., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
21. Ward, R., Scott, N.J.: Cues to mental health from men's facial appearance. J. Res. Pers. **75**, 26–36 (2018)
22. Ward, R., Sreenivas, S., Read, J., Saunders, K.E., Rogers, R.D.: The role of serotonin in personality inference: tryptophan depletion impairs the identification of neuroticism in the face. Psychopharmacology **234**(14), 2139–2147 (2017)
23. Zhang, C.-L., Zhang, H., Wei, X.-S., Wu, J.: Deep bimodal regression for apparent personality analysis. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 311–324. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_25
24. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2921–2929 (2016)