Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition

Yağmur Güçlütürk(✉), Umut Güçlü, Marcel A.J. van Gerven, and Rob van Lier

Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands
{y.gucluturk,u.guclu,m.vangerven,r.vanlier}@donders.ru.nl

Abstract. Here, we develop an audiovisual deep residual network for multimodal apparent personality trait recognition. The network is trained end-to-end for predicting the Big Five personality traits of people from their videos. That is, the network does not require any feature engineering or visual analysis such as face detection, face landmark alignment or facial expression recognition. Recently, the network won the third place in the ChaLearn First Impressions Challenge with a test accuracy of 0.9109.

Keywords: Big five personality traits · Audiovisual · Deep neural network · Deep residual network · Multimodal

1 Introduction

Appearances influence what people think about the personality of other people, even without having any interaction with them. These judgments can be made very quickly - already after 100 ms [35]. Although some studies have shown that people are good at forming accurate first impressions about the personality traits of people after viewing their photographs or videos [4,21], it has also been shown that simply relying on the appearance does not always result in correct first impression judgments [22].

Several characteristics of people varying from clothing to facial expressions, contribute to the first impression judgments about personality [29]. For example, [30] has shown that the photographs of the same person taken with a different facial expression changes the judgments about the person's personality traits such as trustworthiness and extravertedness as well as other perceived characteristics such as attractiveness and intelligence. Furthermore, people are better at guessing other's personality traits if they find them attractive after short encounters with them [18]. The same study also showed that people form more positive first impressions about more attractive people.

Studies of personality prediction generally either deal with correctly recognizing the actual personality traits of people, which can be measured as

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part III, LNCS 9915, pp. 349–358, 2016.
DOI: 10.1007/978-3-319-49409-8_28

self- or acquaintance-reports or apparent personality traits, which are the impressions about the personality of an unfamiliar individual [34]. Below we review the recent work in apparent personality prediction.

Most of the previous work on apparent personality modeling and prediction have been in the domain of paralanguage, i.e. speech, text, prosody, other vocalizations and fillers [34]. Conversations (both text and audio) [19] and speech clips [20,23] were the materials that were most commonly analyzed. In this domain, INTERSPEECH 2012 Speaker Trait Challenge [25] enabled a systematic comparison of computational methods by providing a dataset comprising audio data and extracted features. The competition had three sub-challenges for predicting the Big Five personality traits, likability and pathology of speakers.

Recently, prediction of apparent personality traits from social media content has become a challenge that attracted much attention in the field. For example [6,26] demonstrated that the images that the users "favorite" on Flickr enabled the prediction of both apparent and actual (self-assessed) personality traits of Flickr users. [32] looked at the influence of a large number of physical attributes (e.g. chin length, head size, posture) on people's impressions regarding approachability, youthful-attractiveness and dominance of them. They studied these influences based on people's impressions formed after looking at face photographs. They performed factor analysis to quantify the contribution of physical attributes and used these factors as inputs to a linear neural network to predict impressions. Their predictions were significantly correlated to the actual impression data.

Given that the exact facial expression [30] and the posture [32] of the person in a photograph influences the first impression judgments about that person, as well as the importance of paralinguistic information in impression formation [19], continuous audio-visual data seems to be a suitable medium to study first impressions. In a series of studies using YouTube video blogs (vlogs) [1–3,29], researchers showed that this is indeed the case. Furthermore, [5] showed that audiovisual annotations along with audiovisual cues enabled the best prediction performance for their regression models compared to either using only either one of them.

At the same time, deep neural networks [16,24] in general and deep residual networks [11] in particular have achieved state-of-the-art results in many computer vision tasks. For example, [11] won the first places in the object detection task and the object localization task at the ImageNet Large Scale Visual Recognition Challenge 2015[1] with their seminal work that introduced deep residual networks. Furthermore, deep residual networks have been successfully used in a variety of other computer vision tasks ranging from style transfer [14] and image super-resolution [14] to semantic segmentation [7] and face hallucination [9].

Recently, [33] suggested that deep neural networks can be used for personality trait recognition because of the hierarchical organization of the personality traits [36]. Following this line of reasoning as well as the recent success of deep residual networks, we develop an audiovisual deep residual network for multimodal

[1] http://image-net.org/.

personality trait recognition. The network is trained end-to-end for predicting the apparent Big Five personality traits of people from their videos. That is, the network does not require any feature engineering or visual analysis such as face detection, face landmark alignment or facial expression recognition.

2 Methods

2.1 Architecture

Figure 1 shows an illustration of the network architecture. The network comprises an auditory stream of a 17 layer deep residual network, a visual stream of another 17 layer deep residual network and an audiovisual stream of a fully-connected layer.

The auditory stream and the visual stream are similar to the first 17 layers of the 18 layer deep residual network in [11]. That is, each stream comprises one convolutional layer and eight residual blocks of two convolutional layers. The convolutional layers are followed by batch normalization [13] (all layers), rectified linear units (all layers), max pooling (first layer) and global average pooling (last layer). In the residual blocks that do not change the dimensionality of their inputs, identity shortcut connections are used. In the remaining residual blocks, convolutional shortcut connections are used. In contrast to [11], the number of convolutional kernels are halved.

Similar to [8], the difference between the auditory stream and the visual stream is that inputs, convolutional/pooling kernels and strides of the auditory stream are one-dimensional whereas those of the visual stream are two-dimensional if the number of channels are ignored. That is:

- An $n^2 \times 1 \times 1$ input of the auditory stream corresponds to an $n \times n \times m$ input of the visual stream.
- An $n^2 \times 1 \times m/n^2 \times 1$ convolutional/pooling kernel of the auditory stream corresponds to an $n \times n \times m/n \times n$ convolutional/pooling kernel of the visual stream.
- An $n^2 \times 1$ stride of the auditory stream corresponds to an $n \times n$ stride of the visual stream.

where m is the number of channels.

Outputs of the auditory stream and the visual stream are merged in an audiovisual stream. The audiovisual stream comprises a fully-connected layer. The fully-connected layer is followed by hyperbolic tangent units. Outputs of the audiovisual stream are scaled to [0, 1].

2.2 Training

We used Adam [15] with initial $\alpha = 0.0002$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ and mini-batch size = 32 to train the network by iteratively minimizing the mean absolute error loss function between the target traits and the predicted traits

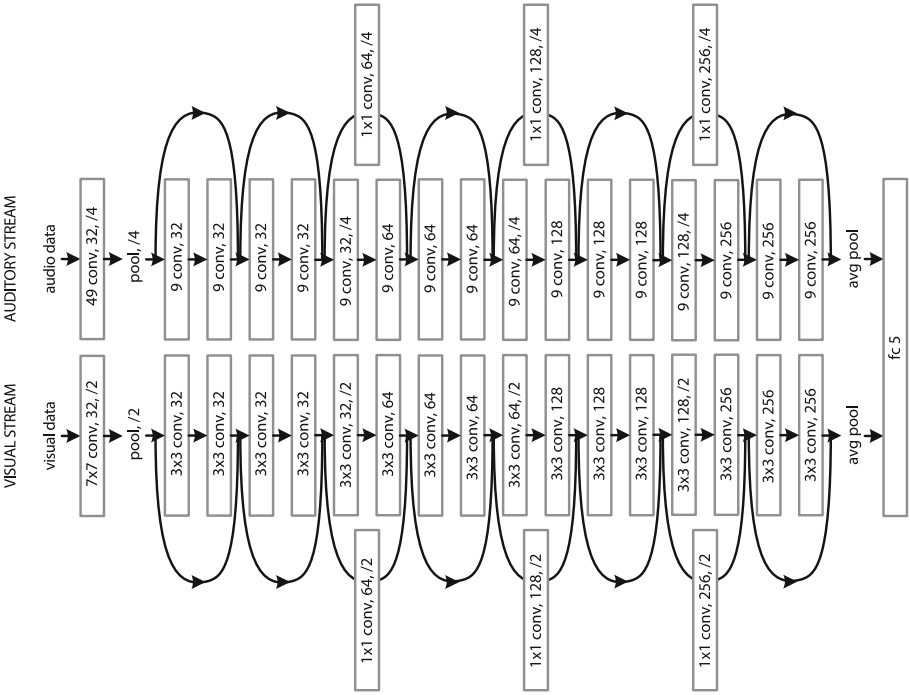

Fig. 1. Illustration of the network architecture.

for 900 epochs. We initialized the biases/weights as in [10] and reduced α by a factor of 10 after every 300 epochs. Each training video clip was processed as follows:

- The audio data and the visual data of the video clip are extracted.
- A random 50176 sample temporal crop of the audio data is fed into the auditory stream. The activities of the penultimate layer of the auditory stream are temporally pooled.
- A random 224 pixels $\times$ 224 pixels spatial crop of a random frame of the visual data is randomly flipped in the left/right direction and fed into the visual stream. The activities of the penultimate layer of the visual stream are spatially pooled.
- The pooled activities of the auditory stream and the visual stream are concatenated and fed into the fully-connected layer.
- The fully-connected layer outputs five continuous prediction values between the range [0, 1] corresponding to each trait for the video clip.

2.3 Validation/Test

Each validation/test video clip was processed as follows:

- The audio data and the visual data of the video clip are extracted.

– The entire audio data are fed into the auditory stream. The activities of the penultimate layer of the auditory stream are temporally pooled (see below note).
– The entire visual data are fed into the visual stream one frame at a time. The activities of the penultimate layer of the visual stream are spatiotemporally pooled (see below note).
– The pooled activities of the auditory stream and the visual stream are concatenated and fed into the fully-connected layer.
– The fully-connected layer outputs five continuous prediction values between the range [0, 1] corresponding to each trait for the video clip.

It should be noted that the network can process video clips of arbitrary sizes since the penultimate layers of the auditory stream and the visual stream are followed by global average pooling.

3 Results

We evaluated the network on the dataset that was released as part of the ChaLearn First Impressions Challenge[2] [17]. The dataset consists of 10000 15-second-long video clips that were drawn from YouTube[3], of which 6000 were used for training, 2000 were used for validation and 2000 were used for test. The video clips were annotated with the Big Five personality traits (i.e. openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) by Amazon Mechanical Turk[4] workers. Each trait was represented with a value between the range [0, 1].

The video clips were preprocessed by temporally resampling the audio data to 16000 Hz as well as spatiotemporally the video data to 456 pixels × 256 pixels and 25 frames per second.

We implemented the network in Chainer [31] with CUDA and cuDNN. Most of the processing took place on a single chip of an Nvidia Tesla K80 GPU accelerator[5]. Processing took approximately 50 ms per training example and 2.7 s per validation/test example on a single chip of an Nvidia Tesla K80 GPU accelerator. Figure 2 shows five validation examples and the corresponding predictions.

Accuracy was defined as 1 - mean absolute error. We report the validation accuracy of the network after 300, 600 and 900 epochs of training (Table 1). Average validation accuracy of the network increased as a function of number of epochs of training with the highest average validation accuracy of 0.9121. We report also the test accuracy of the network after 900 epochs of training, which won the third place in the challenge, and compare it with those of the models that won the first two places in the challenge (Table 2).

[2] http://gesture.chalearn.org.
[3] http://www.youtube.com/.
[4] http://www.mturk.com/.
[5] The implementation is available at https://github.com/yagguc/deep_impression.

Table 1. Validation accuracies of the challenge model after 300, 600 and 900 epochs of training.

Epoch	Validation accuracy					
	Average	Openness	Agreeableness	Conscientiousness	Neuroticism	Extraversion
300	0.906461	0.905451	0.911128	0.902121	0.907886	0.905721
600	0.911929	0.911924	**0.915610**	0.911717	**0.909891**	**0.910503**
900	**0.912132**	**0.911983**	0.915466	**0.913077**	0.909705	0.910429

Table 2. Test accuracies of the models that won the first three places in the challenge.

Rank	Test accuracy					
	Average	Openness	Agreeableness	Conscientiousness	Neuroticism	Extraversion
1 [37]	**0.912968037541**	**0.91237757**	**0.91257098**	**0.91663743**	**0.9099631**	0.91329111
2 [28]	0.912062557634	0.91167725	0.91186694	0.91185413	0.90991703	**0.91499745**
3 (ours)	0.910932616931	0.91108539	0.91019226	0.91377735	0.90890031	0.91070778
			...			
10	0.875888740066	0.87026111	0.88423626	0.87270874	0.87526563	0.87697196

Fig. 2. Example thumbnails of the videos of five people and the corresponding predicted personality traits. Each trait takes a value between [0, 1]. Each color represents a trait. From left to right: Openness, agreeableness, conscientiousness, neuroticism and extraversion.

4 Post Challenge Models

For completeness, we briefly report our preliminary work on two models that we have evaluated after the end of the challenge.

First, we separately fine-tuned the original DNN after 300 epochs of training for each trait. Everything about the fine-tuned DNNs (i.e. architecture, training and validation/test) were the same with the original DNN except for their fully-connected layers that output one value rather than five values.

Table 3. Validation accuracies of the challenge model and the post challenge models.

Model	Validation accuracy					
	Average	Openness	Agreeableness	Conscientiousness	Neuroticism	Extraversion
DNN	0.912132	**0.911983**	0.915466	0.913077	**0.909705**	0.910429
5 × DNN	0.911987	0.911522	0.915413	0.913211	0.909062	0.910727
DNN + RNN	**0.912158**	0.911676	**0.915761**	**0.913300**	0.909056	**0.910996**

Second, we trained a recurrent neural network (RNN) on top of the original network. The RNN comprised two layers of 512 long short-term memory units [12] and one layer of five linear units. At each time point, the RNN took as input the layer 5 features of a second-long video clip and the output of the RNN was the predicted traits. Dropout [27] was used to regularize the hidden layers.

We used Adam to train the model by iteratively minimizing the mean absolute error loss function between the target traits and the predicted traits at each time point. Backpropagation was truncated after every 15 time points. Once the model was trained, the predicted traits were averaged over the entire video clip.

Table 3 shows the validation accuracy of the post challenge models. While the post challenge models failed to outperform the challenge model to a large extent, we strongly believe that variants thereof have the potential to do so and will be the subject matter of future work.

5 Conclusion

In this study, we presented our approach and results that won the third place in the ChaLearn First Impressions Challenge. Summarizing, we developed and trained an audiovisual deep residual network for predicting the apparent personality traits of people in an end-to-end manner. This approach enabled us to obtain very high performance for all traits while exploiting the similarities between the organization of the personality traits and the deep neural networks in terms of the hierarchical organization and circumventing extensive analyses for identifying/designing relevant features for the task of apparent personality traits prediction. Our results demonstrate the potential of deep neural networks in the field of automatic (perceived) personality prediction. Future work will focus on the extensions of the current work with recurrent neural networks and language models as well as identifying the factors that drive first impressions.

References

1. Biel, J.I., Aran, O., Gatica-Pere, D.: You are known by how you vlog: personality impressions and nonverbal behavior in youtube. In: International Conference on Weblogs and Social Media (2011)
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). http://dx.doi.org/10.1109/TMM.2012.2225032

3. Biel, J.I., Teijeiro-Mosquera, L., Gatica-Perez, D.: FaceTube. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction. Association for Computing Machinery (ACM) (2012). http://dx.doi.org/10.1145/2388676.2388689
4. Borkenau, P., Liebler, A.: Trait inferences: surces of validity at zero acquaintance. J. Pers. Soci. Psychol. **62**(4), 645–657 (1992). http://dx.doi.org/10.1037/0022-3514.62.4.645
5. Celiktutan, O., Gunes, H.: Automatic prediction of impressions in time and across varying context: personality, attractiveness and likeability. In: IEEE Transaction on Affective Computing, p. 1 (2016). http://dx.doi.org/10.1109/TAFFC.2015.2513401
6. Cristani, M., Vinciarelli, A., Segalin, C., Perina, A.: Unveiling the multimedia unconscious. In: Proceedings of the 21st ACM International Conference on Multimedia. Association for Computing Machinery (ACM) (2013). http://dx.doi.org/10.1145/2502081.2502280
7. Dai, J., He, K., Sun, J.: Instance-aware semantic segmentation via multi-task network cascades. CoRR abs/1512.04412 (2015)
8. Güçlü, U., Thielen, J., Hanke, M., van Gerven, M.A.J.: Brains on beats. CoRR abs/1606.02627 (2016)
9. Güçlütürk, Y., Güçlü, U., van Lier, R., van Gerven, M.A.J.: Convolutional sketch inversion. CoRR abs/1606.03073 (2016)
10. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. CoRR abs/1406.4729 (2014)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). http://dx.doi.org/10.1162/neco.1997.9.8.1735
13. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR abs/1502.03167 (2015)
14. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. CoRR abs/1603.08155 (2016)
15. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014)
16. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015). http://dx.doi.org/10.1038/nature14539
17. Lopez, V.P., Chen, B., Places, A., Oliu, M., Corneanu, C., Baro, X., Escalante, H.J., Guyon, I., Escalera, S.: ChaLearn LaP 2016: first round challenge on first impressions - dataset and results. In: ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop proceedings. Springer Science + Business Media, Berlin (2016, in press)
18. Lorenzo, G.L., Biesanz, J.C., Human, L.J.: What is beautiful is good and more accurately understood: physical attractiveness and accuracy in first impressions of personality. Psychol. Sci. **21**(12), 1777–1782 (2010). http://dx.doi.org/10.1177/0956797610388048
19. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using linguistic cues for the automatic recognition of personality in conversation and text. J. Artif. Intell. Res. **30**(1), 457–500 (2007). http://dl.acm.org/citation.cfm?id=1622637.1622649
20. Mohammadi, G., Vinciarelli, A.: Automatic personality perception: prediction of trait attribution based on prosodic features extended abstract. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). Institute of Electrical & Electronics Engineers (IEEE), September 2015. http://dx.doi.org/10.1109/ACII.2015.7344614

21. Naumann, L.P., Vazire, S., Rentfrow, P.J., Gosling, S.D.: Personality judgments based on physical appearance. Pers. Soc. Psychol. Bull. **35**(12), 1661–1671 (2009). http://dx.doi.org/10.1177/0146167209346309
22. Olivola, C.Y., Todorov, A.: Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. J. Exp. Soc. Psychol. **46**(2), 315–324 (2010). http://dx.doi.org/10.1016/j.jesp.2009.12.002
23. Polzehl, T., Moller, S., Metze, F.: Automatically assessing personality from speech. In: 2010 IEEE Fourth International Conference on Semantic Computing. Institute of Electrical & Electronics Engineers (IEEE), September 2010. http://dx.doi.org/10.1109/ICSC.2010.41
24. Schmidhuber, J.: Deep learning in neural networks: an overview. Neural Netw. **61**, 85–117 (2015). http://dx.doi.org/10.1016/j.neunet.2014.09.003
25. Schuller, B., Steidl, S., Batliner, A., Nöth, E., Vinciarelli, A., Burkhardt, F., van Son, R., Weninger, F., Eyben, F., Bocklet, T., Mohammadi, G., Weiss, B.: A survey on perceived speaker traits: personality, likability, pathology, and the first challenge. Comput. Speech Lang. **29**(1), 100–131 (2015). http://dx.doi.org/10.1016/j.csl.2014.08.003
26. Segalin, C., Perina, A., Cristani, M., Vinciarelli, A.: The pictures we like are our image: continuous mapping of favorite pictures into self-assessed and attributed personality traits. In: IEEE Transactions on Affective Computing, p. 1 (2016). http://dx.doi.org/10.1109/TAFFC.2016.2516994
27. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
28. Subramaniam, A., Patel, V., Mishra, A., Balasubramanian, P., Mittal, A.: Bimodal first impressions recognition using temporally ordered deep audio and stochastic visual features. In: ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop proceedings. Springer Science + Business Media, Berlin (2016, in press)
29. Teijeiro-Mosquera, L., Biel, J.I., Alba-Castro, J.L., Gatica-Perez, D.: What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube. IEEE Trans. Affective Comput. **6**(2), 193–205 (2015). http://dx.doi.org/10.1109/TAFFC.2014.2370044
30. Todorov, A., Porter, J.M.: Misleading first impressions: different for different facial images of the same person. Psychol. Sci. **25**(7), 1404–1417 (2014). http://dx.doi.org/10.1177/0956797614532474
31. Tokui, S., Oono, K., Hido, S., Clayton, J.: Chainer: a next-generation open source framework for deep learning. In: Workshop on Machine Learning Systems at Neural Information Processing Systems (2015)
32. Vernon, R.J.W., Sutherland, C.A.M., Young, A.W., Hartley, T.: Modeling first impressions from highly variable facial images. Proc. Natl. Acad. Sci. **111**(32), E3353–E3361 (2014). http://dx.doi.org/10.1073/pnas.1409860111
33. Vinciarelli, A., Mohammadi, G.: More personality in personality computing. IEEE Trans. Affect. Comput. **5**(3), 297–300 (2014). http://dx.doi.org/10.1109/TAFFC.2014.2341252
34. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. IEEE Trans. Affect. Comput. **5**(3), 273–291 (2014). http://dx.doi.org/10.1109/TAFFC.2014.2330816
35. Willis, J., Todorov, A.: First impressions: making up your mind after a 100-ms exposure to a face. Psychol. Sci. **17**(7), 592–598 (2006). http://dx.doi.org/10.1111/j.1467-9280.2006.01750.x

36. Wright, A.G.: Current directions in personality science and the potential for advances through computing. IEEE Trans. Affect. Comput. **5**(3), 292–296 (2014). http://dx.doi.org/10.1109/TAFFC.2014.2332331
37. Zhang, C.L., Zhang, H., Wei, X.S., Wu, J.: Deep bimodal regression for apparent personality analysis. In: ChaLearn Looking at People Workshop on Apparent Personality Analysis, ECCV Workshop proceedings. Springer Science + Business Media (2016, in press)