Combining Deep Facial and Ambient Features for First Impression Estimation

Furkan Gürpınar[1], Heysem Kaya[2], Albert Ali Salah[3]

[1]Program of Computational Science and Engineering, Boğaziçi University, Bebek, Istanbul, Turkey
furkan.gurpinar@boun.edu.tr
[2]Department of Computer Engineering, Namık Kemal University, Çorlu, Tekirdağ, Turkey
hkaya@nku.edu.tr
[3]Department of Computer Engineering, Boğaziçi University, Bebek, Istanbul, Turkey
salah@boun.edu.tr

Abstract. First impressions influence the behavior of people towards a newly encountered person or a human-like agent. Apart from the physical characteristics of the encountered face, the emotional expressions displayed on it, as well as ambient information affect these impressions. In this work, we propose an approach to predict the first impressions people will have for a given video depicting a face within a context. We employ pre-trained Deep Convolutional Neural Networks to extract facial expressions, as well as ambient information. After video modeling, visual features that represent facial expression and scene are combined and fed to Kernel Extreme Learning Machine regressor. The proposed system is evaluated on the ChaLearn Challenge Dataset on First Impression Recognition, where the classification target is the "Big Five" personality trait labels for each video. Our system achieved an accuracy of 90.94% on the sequestered test set, 0.36% points below the top system in the competition.

Keywords: personality traits, first impression, deep learning, ELM

1 Introduction and Related Work

It is not possible to judge the personality of a person by a mere glimpse of the face, but people attribute apparent personality traits for a face they newly encounter, in a stereotypical way, and with remarkable consistency [1]. In this work, we tackle the problem of predicting the apparent personality using the data and protocol from the ChaLearn Looking at People 2016 First Impression Challenge [2].

It is not surprising that emotional expressions influence the attribution of personality traits. It is more likely for a smiling person to be perceived as more trustworthy, and friendly. Todorov et al. convincingly argued that rapid, unreflective trait inferences from faces can influence consequential decisions [3]. This

is why people do not typically use frowning or angry pictures in their resumés. Also the context of the image can affect the perception of the face. In our proposed approach, we estimate emotional facial expressions, as well as cues from the context of the face to predict first impressions.

Before describing the followed approach, we provide a brief literature review on automatic personality trait recognition. In the past, various approaches have been used for recognizing apparent personality traits from different modalities such as audio [4, 5], text [6–8] and visual information [9, 10]. As in other recognition problems, multimodal systems are also investigated to improve robustness of prediction [11–14]. These works aim to estimate personality traits from given input. In psychology, personality is often assessed by running a "Big Five" questionnaire that measures Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) [15]. Apparent personality is also frequently assessed in these five dimensions.

In their work, Borkenau and Liebler used the Brunswik's lens model and categorized the particular cues that may communicate a certain personality [16]. They included a large number of indicators such as overall impression variables (e.g. estimated age, masculinity, attractiveness), acoustic variables (e.g. softness of voice, pleasantness, clarity), static visual variables (e.g. appearance, makeup, garments, thin lips, hair style, facial expression), and dynamic visual variables (e.g. movement speed, hand movements, walking style). In order to assess the personality trait attributions, they measured "validity," which indicates the correlation between self-ratings of personality and ratings by strangers or acquaintances. The Brunswik's lens model looks at cues used for perceived traits, and links some of these cues to actual traits by assessing their ecological validity [17]. It is a useful conceptualization, also used in approaches to personality computing [18].

According to the literature, faces are a rich source of cues for apparent personality attribution, related to stereotype judgments. For an automatic analysis system, the first steps of a visual face analysis pipeline are face detection [19, 20] and facial landmark localization [21–23]. Face alignment (or registration) is an important step, as all further processing depends on its accuracy. Recent deep neural network approaches are known to be more resistant to registration errors.

Face alignment is followed by visual feature extraction, which can include image-level appearance descriptors such as Local Binary Patterns (LBP) [24], Histogram of Oriented Gradients (HOG) [25], Scale-invariant Feature Transform (SIFT) [26], video-level descriptors such as Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [27] and Local Phase Quantization (LPQ)-TOP [28], or geometric information [9, 10].

Deep learning based approaches have achieved state-of-the-art results in human behavior analysis. These approaches, when trained with large datasets, can provide representations that are very robust to variations exhibited in the data. Deep learning has been successfully applied to many tasks related to computer vision such as object recognition [29, 30], face recognition [31], emotion recognition [32] and age estimation [33–37]. Moreover, deep representations of images

are often usable for multiple tasks, enabling transfer learning from pre-trained models. The disadvantages are the relatively high computational requirements for training such systems, the large amount of training data required, and (relatively) poor temporal extension to video processing.

In recent approaches to personality impressions classification, Support Vector Machines (SVM) [38] have been widely used [5, 8, 12, 14]. Recently, a learning approach called Extreme Learning Machines (ELM) that is similar to SVMs but providing faster learning schemes has become popular [39]. The use of ELM's name is debated in the literature, because of its strong resemblance to earlier methods. We continue to use it in this work for convenience. The approach has been shown to provide good performance in a number of applications including face recognition [40, 41], emotion recognition [42, 43], and smile detection [44].

Given the success of deep learning approaches and the speed of ELM, we propose to use a fusion of deep face and scene features, followed by regularized regression with a kernel ELM classifier. The main contribution of this work is the effective combination of emotion related and ambient features that are efficiently extracted from pre-trained/fine-tuned Deep Convolutional Neural Network (DCNN) models. Our method is illustrated in a simplified flowchart in Figure 1.

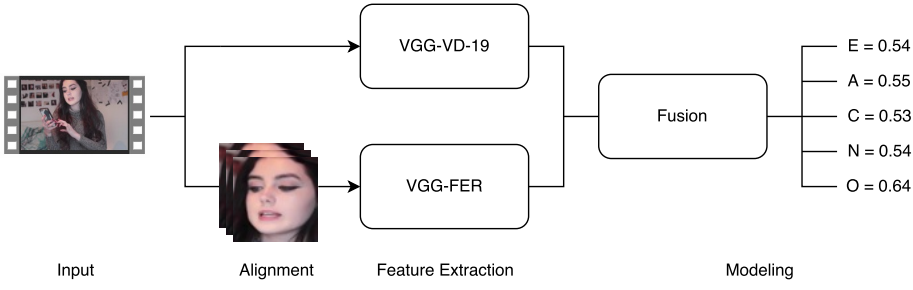


Fig. 1. Flowchart of the proposed method.

The remainder of this paper is organized as follows. In the next section we provide background and details on the methodology. Then in Section 3, we present the experimental results, followed by implementation details in Section 4. Finally, Section 5 concludes with future directions.

2 Methodology

Our proposed approach evaluates a short video clip that contains a single person, and outputs an estimate of apparent personality traits in the five dimensions mentioned earlier. In this section, we describe the three main steps of our pipeline, namely, face alignment, feature extraction, and modeling.

2.1 Face Alignment

For detecting and aligning faces from the videos, we use Xiong and de la Torre's Supervised Descent Method (SDM), also known as IntraFace [21]. This approach locates 49 landmarks on the face. After the landmarks are located, we estimate the roll angle of the face from the eye corner locations and rotate the image to rectify the face. We then add a margin of 20% interocular distance around the outer landmarks to compute a loose bounding box from which we crop facial images. After the face is cropped, it is resized to 64×64 pixels, and registered as a new frame. Frames from a sample input video and the corresponding aligned face images are shown in Figure 2.

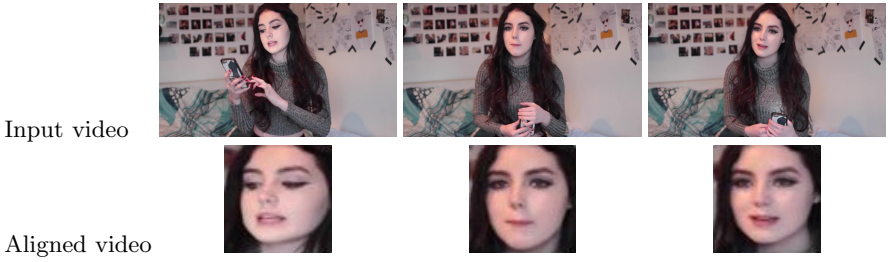


Fig. 2. Face alignment example.

2.2 Feature Extraction

We extract facial features that are summarized over an entire video segment, and scene features from the first image of each video. The assumption is that videos do not stretch over multiple shots.

Face Features: After aligning the faces, we extract image-level deep features from a network that is trained for facial emotion recognition. The training of this network is explained in more detail in Section 2.3. For comparison, we also extract features from the original VGG-Face network that was trained for face recognition [31]. For both networks, we use the response of the 33^{rd} layer of the 37-layer architecture, which is the lowest-level 4096-dimensional descriptor.

We compare deep features with traditional appearance descriptors and geometric information that is shown to be effective in emotion recognition [45]. We report the cross validation accuracy of each approach in Section 3.2.

Video Features: After extracting frame-level features from each registered face, we summarize the videos by computing functional statistics of each dimension over time. The functionals include mean, standard deviation, offset, slope, and curvature. Offset and slope are calculated from the first order polynomial fit

to each feature contour, while curvature is the leading coefficient of the second order polynomial. An empirical comparison of the individual functionals is given in Section 3.2.

Scene Features: In order to use ambient information in the images to our advantage, we extract features using the VGG-19 network [30], which is trained for an object recognition task on the ILSVRC 2012 dataset. Similar to face features, we use the 4096-dimensional feature from the 39^{th} layer of the 43-layer architecture, hence we obtain a description of the overall image that contains both the face and the scene, which we combine with face features using feature-level fusion.

2.3 CNN Fine Tuning

We start with the VGG-Face network [31], changing the final layer (originally a 2622-dimensional recognition layer), to a 7-dimensional emotion recognition layer, where the weights are initialized randomly. We fine-tune this network with the softmax loss function using around 30,000 training images from the FER-2013 dataset [46]. We choose an initial learning rate of 0.0001, a momentum of 0.9 and a batch size of 64. We train the model for 5 epochs, and we show the validation set performance for each epoch in Figure 3.

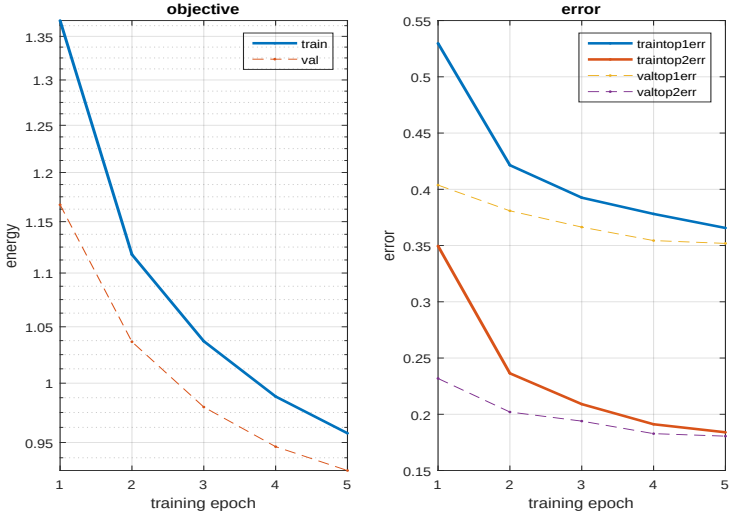


Fig. 3. Fine tuning the VGG-Face network on the FER-2013 public test set. The figure on the left shows the softmax loss, whereas the figure on the right shows the top-1 and top-2 classification errors.

2.4 Regression with Kernel ELM

In order to model personality traits from visual features, we used kernel ELM, due to the learning speed and accuracy of the algorithm. In the following paragraphs, we briefly explain the learning strategy of ELM.

ELM proposes a simple and robust learning algorithm for single-hidden layer feedforward networks. The input layer's bias and weights are initialized randomly to obtain the output of the second (hidden) layer. The bias and weights of the second layer are calculated by a simple generalized inverse operation of the hidden layer output matrix.

ELM tries to find the mapping between the hidden node output matrix $\mathbf{H} \in \mathbb{R}^{N \times h}$ and the label vector $\mathbf{T} \in \mathbb{R}^{N \times 1}$ where N and h denote the number of samples and the hidden neurons, respectively. The set of output weights $\beta \in \mathbb{R}^{h \times 1}$ is calculated by the least squares solution of the set of linear equations $\mathbf{H}\beta = \mathbf{T}$, as:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{1}$$

where $\mathbf{H}^{\dagger}$ denotes the Moore-Penrose generalized inverse [47] that minimizes the L_2 norms of $||\mathbf{H}\beta - \mathbf{T}||$ and $||\beta||$ simultaneously.

To increase the robustness and the generalization capability of ELM, a regularization coefficient $\mathbf{C}$ is included in the optimization procedure. Therefore, given a kernel $\mathbf{K}$, the set of weights is learned as follows:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}. \tag{2}$$

In order to prevent parameter overfitting, we use the linear kernel $\mathbf{K}(x, y) = x^T y$, where x and y are the original feature vectors after min-max normalization of each dimension among the training samples. With this approach, the only parameter of our model is the regularization coefficient C, which we optimize with a 5-fold subject independent cross-validation on the training set. In Section 3.2, we report the average score of each fold with the selected parameter.

3 Experiments

3.1 Challenge and Corpus

The "ChaLearn LAP Apparent Personality Analysis: First Impressions" challenge consists of 10,000 clips collected from 5,563 YouTube videos, where the poses are more or less frontal, but the resolution, lighting and background conditions are not controlled, hence providing a dataset with in-the-wild conditions. Each clip in the training set is labeled for the Big Five personality traits. Basic statistics of the dataset partitions are provided in Table 1. The detailed information on the challenge and corpus can be found in [2].

Table 1. Dataset summary

	Train	Val	Test
#Clips	6,000	2,000	2,000
#YouTube videos	2,624	1,484	1,455
#Given frames	2.56M	0.86M	0.86M
#Detected frames	2.45M	0.82M	0.82M

Performance Evaluation: The performance score in this challenge is the Mean Absolute Error subtracted from 1, which is formulated as follows:

$$1 - \sum_{i}^{N} \frac{|\hat{y}_i - y_i|}{N}, \tag{3}$$

where N is the number of samples, $\hat{y}$ is the predicted label and y is the true label ($0 \leq y \leq 1$). This score is then averaged over five tasks. This means the final score varies between 0 (worst case) and 1 (best case).

3.2 Experimental Results

In this section, we report the regression performance of various visual descriptors. Tables 2 and 3 summarize the performances of the different systems with 5-fold subject-independent cross-validation on the training set.

We first look at the performance of individual functionals, which are described in Section 2.2. As can be seen in Table 2, the combination of mean, standard deviation, and offset features works well, and the mean by itself is the most informative functional.

Table 2. Functional statistics with deep face features.

Feature	**Mean**	**Extr.**	**Agre.**	**Cons.**	**Neur.**	**Open.**
Mean	0.900	0.906	0.902	0.897	0.894	0.902
Std	0.883	0.891	0.881	0.876	0.880	0.886
Curvature	0.880	0.876	0.891	0.874	0.876	0.882
Slope	0.880	0.876	0.892	0.874	0.876	0.882
Offset	0.899	0.904	0.901	0.895	0.893	0.901
Fusion of all 5	0.902	0.908	0.903	0.898	**0.898**	0.904
Mean + Std + Offset	**0.902**	**0.909**	**0.903**	**0.899**	0.897	**0.904**

We evaluate a set of features with different dimensionalities individually. Geometric features (GEO), LPQ-TOP, LBP-TOP, and different deep neural network features were individually tested. Table 3 summarizes the results, and gives the dimensionality of each selected feature set. We observe that features from the

deep face model fine tuned on the FER emotion corpus provide higher performances compared to both original deep features and hand-crafted visual features. Combining these features with ambient (scene) information further improves the prediction performance.

Table 3. Regression performance with various visual descriptors

ID	Feature	Dim.	Mean	Extr.	Agre.	Cons.	Neur.	Open.
1	GEO	115	0.892	0.896	0.896	0.883	0.888	0.896
2	LPQ-TOP	12288	0.901	0.904	0.901	0.898	0.899	0.903
3	LBP-TOP	5568	0.900	0.903	0.900	0.895	0.897	0.902
4	LGBP-TOP	100224	0.903	0.907	0.902	0.900	**0.901**	0.905
5	VGG-19	4096	0.890	0.886	0.895	0.892	0.884	0.894
6	Caffe-Alex	4096	0.890	0.887	0.895	0.890	0.885	0.894
7	VGGFace	12288	0.901	0.907	0.901	0.898	0.896	0.903
8	VGGFace+FER	12288	0.902	0.909	0.903	0.899	0.897	0.904
9	Fusion (5&8)	16384	**0.904**	**0.909**	**0.904**	**0.902**	0.899	**0.907**

The best fusion system (ID 9 in Table 3) gives a test set mean accuracy of 0.9094, which ranks the fifth in the official competition. Considering the obtained test set performance in comparison to other competitors' accuracies (see Table 4), we observe that the performances are around 0.90-0.91 in general. The top accuracy is 0.9130, while the top six teams' accuracies are higher than 0.9.

Table 4. Final ranking on the test set

Rank	Team	Accuracy
1	NJU-LAMDA	0.9130
2	evolgen	0.9121
3	DCC	0.9109
4	ucas	0.9098
5	**BU-NKU (ours)**	0.9094
6	pandora	0.9063
7	Pilab	0.8936
8	Kaizoku	0.8826
9	ITU_SiMiT	0.8815
10	sp	0.8759

We show the estimations of our system during cross validation in Figures 4 and 5. The results in Figure 4 show how precisely our system can estimate the personality traits under various imaging conditions. Figure 5 shows that examples with labels very close to 0 or 1 tend to have higher error, which might

be due to the approximately normal distribution of training labels with mean values around 0.5.

4 Implementation Details

The whole system is implemented in MATLAB R2015b on a 64-bit Windows 10 PC with 32GB RAM and an Intel i7-6700 CPU. For fine-tuning and feature extraction with CNNs, the MatConvNet library [48] has been used with GPU parallelization, using an NVidia GeForce GTX 970 GPU. Time spent on important parts of the pipeline is summarized in Table 5.

Table 5. Time requirement for each step of the pipeline

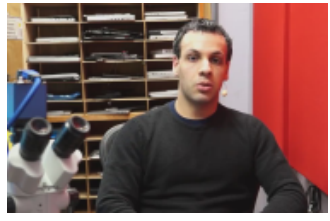
Task	**Time**	**Unit**
Face det. & alignment	0.17s	per image
Feature extraction (w/o GPU)	0.24s	per image
Feature extraction (with GPU)	0.03s	per image
Functional encoding	3s	per video
Kernel ELM training	0.37s	for train set
Kernel ELM testing	10^{-5}s	per video
Total	98s	per video

5 Conclusions

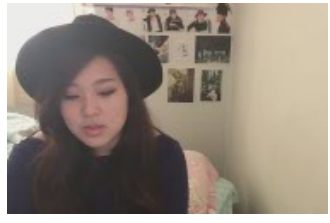
In this paper, we proposed to use transfer learning in order to estimate the personality trait perceptions during first impressions. We use deep convolutional neural networks (DCNN) that are originally trained for other tasks such as face, object, and emotion recognition, and we employ their features directly. Hence, we show the feasibility of deep transfer learning for this task.

Combining two sets of DCNN features that carry facial expression and ambient information, we achieve better results compared to each of these approaches, as well as compared to other hand-crafted visual features. In this work, we did not make use of the audio modality, which was shown to be beneficial in earlier works. Audio-based and multimodal analyses constitute our future work. In this work, video modeling is carried out using simple statistical functionals. This approach is fast and shown to be accurate. For future works, a wider set of functionals will be investigated.

	True	Pred.	True	Pred.
Open.	0.64	0.68	0.58	0.59
Cons.	0.59	0.59	0.42	0.40
Extr.	0.65	0.64	0.47	0.49
Agre.	0.63	0.64	0.54	0.53
Neur.	0.60	0.62	0.47	0.49

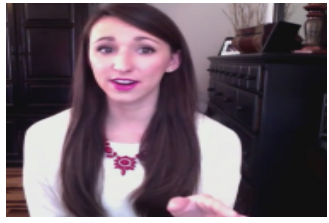


	True	Pred.	True	Pred.
Open.	0.57	0.54	0.69	0.70
Cons.	0.70	0.71	0.61	0.58
Extr.	0.38	0.40	0.64	0.65
Agre.	0.60	0.57	0.62	0.64
Neur.	0.54	0.56	0.65	0.64

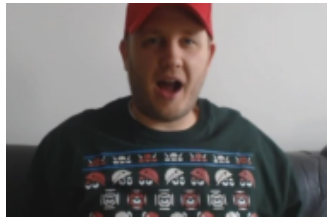
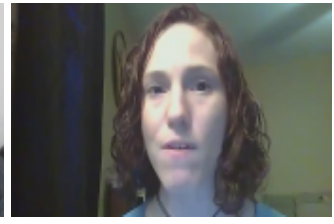


	True	Pred.	True	Pred.
Open.	0.56	0.54	0.46	0.45
Cons.	0.50	0.47	0.51	0.53
Extr.	0.46	0.47	0.36	0.34
Agre.	0.54	0.53	0.40	0.45
Neur.	0.48	0.47	0.36	0.37

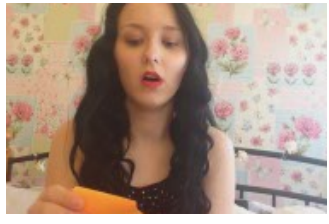
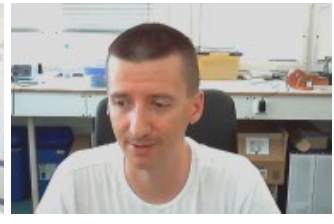
Fig. 4. Six examples from the training set where our approach produced good estimations for the traits. For each example, the first colums shows the ground truth (True), and the second column shows the estimation of the model (Pred.)




	True	Pred.	True	Pred.
Open.	0.27	0.69	0.71	0.48
Cons.	0.21	0.75	0.66	0.40
Extr.	0.17	0.62	0.66	0.35
Agre.	0.11	0.64	0.76	0.47
Neur.	0.14	0.65	0.75	0.45

	True	Pred.	True	Pred.
Open.	0.91	0.50	0.13	0.52
Cons.	0.67	0.47	0.06	0.41
Extr.	0.72	0.39	0.02	0.41
Agre.	0.74	0.50	0.02	0.48
Neur.	0.77	0.46	0.09	0.44

	True	Pred.	True	Pred.
Open.	0.17	0.62	0.16	0.48
Cons.	0.22	0.51	0.35	0.58
Extr.	0.12	0.52	0.10	0.39
Agre.	0.23	0.56	0.22	0.55
Neur.	0.20	0.52	0.23	0.46

Fig. 5. Examples from the training set where our approach produced poor estimations for the traits. For each example, the first colums shows the ground truth (True), and the second column shows the estimation of the model (Pred.)

References

1. Cuddy, A.J., Fiske, S.T., Glick, P.: Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. Advances in experimental social psychology **40** (2008) 61–149
2. 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. (2016)
3. Todorov, A., Mandisodza, A.N., Goren, A., Hall, C.C.: Inferences of competence from faces predict election outcomes. Science **308**(5728) (2005) 1623–1626
4. Valente, F., Kim, S., Motlicek, P.: Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus. In: INTERSPEECH. (2012) 1183–1186
5. Madzlan, N., Han, J., Bonin, F., Campbell, N.: Towards automatic recognition of attitudes: Prosodic analysis of video blogs. Speech Prosody, Dublin, Ireland (2014) 91–94
6. Alam, F., Stepanov, E.A., Riccardi, G.: Personality traits recognition on social network-facebook. WCPR (ICWSM-13), Cambridge, MA, USA (2013)
7. Nowson, S., Gill, A.J.: Look! who's talking?: Projection of extraversion across different social contexts. In: Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition, ACM (2014) 23–26
8. Gievska, S., Koroveshovski, K.: The impact of affective verbal content on predicting personality impressions in youtube videos. In: Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition, ACM (2014) 19–22
9. Fernando, T., et al.: Persons' personality traits recognition using machine learning algorithms and image processing techniques. Advances in Computer Science: an International Journal **5**(1) (2016) 40–44
10. Qin, R., Gao, W., Xu, H., Hu, Z.: Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face. arXiv preprint arXiv:1604.07499 (2016)
11. Sarkar, C., Bhatia, S., Agarwal, A., Li, J.: Feature analysis for computational personality recognition using youtube personality data set. In: Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition, ACM (2014) 11–14
12. Alam, F., Riccardi, G.: Predicting personality traits using multimodal information. In: Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition, ACM (2014) 15–18
13. Farnadi, G., Sushmita, S., Sitaraman, G., Ton, N., De Cock, M., Davalos, S.: A multivariate regression approach to personality impression recognition of vloggers. In: Proceedings of the 2014 ACM Multimedia Workshop on Computational Personality Recognition, ACM (2014) 1–6
14. Sidorov, M., Ultes, S., Schmitt, A.: Automatic recognition of personality traits: A multimodal approach. In: Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge, ACM (2014) 11–15
15. Gosling, S.D., Rentfrow, P.J., Swann, W.B.: A very brief measure of the big-five personality domains. Journal of Research in personality **37**(6) (2003) 504–528
16. Borkenau, P., Liebler, A.: Trait inferences: Sources of validity at zero acquaintance. Journal of Personality and Social Psychology **62**(4) (1992) 645

17. Zebrowitz, L.A., Collins, M.A.: Accurate social perception at zero acquaintance: The affordances of a gibsonian approach. Personality and social psychology review **1**(3) (1997) 204–223
18. Vinciarelli, A., Mohammadi, G.: A survey of personality computing. IEEE Transactions on Affective Computing **5**(3) (2014) 273–291
19. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision **57**(2) (2004) 137–154
20. Mathias, M., Benenson, R., Pedersoli, M., Van Gool, L.: Face detection without bells and whistles. Computer Vision–ECCV 2014 (2014) 720–735
21. Xiong, X., De la Torre, F.: Supervised Descent Method and Its Application to Face Alignment. In: IEEE Conference on Computer Vision and Pattern Recognition. (2013) 532–539
22. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1685–1692
23. Xiong, X., De la Torre, F.: Global supervised descent method. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2664–2673
24. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(7) (2002) 971–987
25. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 1., IEEE (2005) 886–893
26. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
27. Almaev, T.R., Valstar, M.F.: Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE (2013) 356–361
28. Jiang, B., Valstar, M.F., Pantic, M.: Action unit detection using sparse appearance descriptors in space-time video volumes. In: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE (2011) 314–321
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
31. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: British Machine Vision Conference. (2015)
32. Kim, B.K., Lee, H., Roh, J., Lee, S.Y.: Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM (2015) 427–434
33. Rothe, R., Timofte, R., Gool, L.: Dex: Deep expectation of apparent age from a single image. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 10–15
34. Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., Chen, X.: Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 16–24

35. Zhu, Y., Li, Y., Mu, G., Guo, G.: A study on apparent age estimation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 25–31
36. Escalera, S., Torres, M., Martinez, B., Baro, X., Jair Escalante, H., Guyon, I., Tzimiropoulos, G., Corneou, C., Oliu, M., Bagheri, M.A., Valstar, M.: Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (June 2016) 1–8
37. Gürpınar, F., Kaya, H., Dibeklioğlu, H., Salah, A.A.: Kernel ELM and CNN Based Facial Age Estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Las Vegas, Nevada, USA (June 2016) 80–86
38. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3) (1995) 273–297
39. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **42**(2) (2012) 513–529
40. Zong, W., Huang, G.B.: Face recognition based on extreme learning machine. Neurocomputing **74**(16) (2011) 2541–2551
41. Mohammed, A.A., Minhas, R., Wu, Q.J., Sid-Ahmed, M.A.: Human face recognition based on multidimensional pca and extreme learning machine. Pattern Recognition **44**(10) (2011) 2588–2597
42. Utama, P., Ajie, H., et al.: A framework of human emotion recognition using extreme learning machine. In: Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of, IEEE (2014) 315–320
43. Kaya, H., Karpov, A.A., Salah, A.A. In: Robust Acoustic Emotion Recognition based on Cascaded Normalization and Extreme Learning Machines. Volume 9719 of Lecture Notes in Computer Science. Springer (2016) 115–123
44. An, L., Yang, S., Bhanu, B.: Efficient smile detection by extreme learning machine. Neurocomputing **149** (2015) 354–363
45. Kaya, H., Gürpınar, F., Afshar, S., Salah, A.A.: Contrasting and combining least squares based learners for emotion recognition in the wild. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM (2015) 459–466
46. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. In: International Conference on Neural Information Processing, Springer (2013) 117–124
47. Rao, C.R., Mitra, S.K.: Generalized inverse of matrices and its applications. Volume 7. Wiley New York (1971)
48. Vedaldi, A., Lenc, K.: MatConvNet – Convolutional Neural Networks for MATLAB. (2015)