Automatic video analysis of personality traits

Daniel Helm*
Supervised by: Martin Kampel†

Institute of Visual Computing & Human-Centered Technology
Vienna University of Technology
Vienna / Austria

Abstract

The first impression of a person can decide about a positive or negative outcome in different situations. The human brain is able to get a picture of the counterpart's personality at short notice. The main aim of this paper is to comprehend how an automated system can be built to predict the Big-Five personality traits of a person. Therefore, a Convolutional Neural Network is trained on visual data features extracted from short video sequences. This paper investigates how various pre-processing methods, like face-extraction and data-augmentation, influence the predicted personality confidences. Furthermore, it explores different deep learning techniques e.g. regularization in order to improve the video-based predictions of the Big-Five personality traits. Moreover, this paper points out the complexity of developing and training a Convolutional Neural Network based system to solve a regression task. Finally, the results derived in this paper are compared to those reported by the winning teams of the First Impressions Challenge 2016 organized by ChaLearn Looking At People, published at the European Conference of Computer Vision 2016.

Keywords: first impressions analysis, apparent personality analysis, deep regression learning, convolutional neural networks, big-five personality traits, deep neural network

1 Introduction

Today the first impression of a person can have significant relevance in daily life. It can decide about a positive or negative result in many situations, like job interviews or business sales hearings. For this reason, many people try to improve their appearance in specific situations. Today, the most common methods for achieving this are communication or presentation trainings with rhetoric coaches. This paper evaluates the possibility of detecting someones impression on others by using a video analysis algorithm.

In the research field of psychology, currently, one of the dominant personality paradigms is the Big-Five personality model [11]. This model describes the personality by means of five dimensions: Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N) and Openness for Experience (O). Each individual dimension of the model is represented as a continuous number between 0 and 1 and all five dimensions are called the Big-Five confidences. People who are curious and inventive count to the dimension Openness to Experience. Conscientiousness corresponds to persons who are strongly organized and efficient. Agreeable humans tend to be friendly or compassionate, whereas more sensitive or nervous persons incline to be neurotic. Extraversion corresponds to people who are energetic and strongly communicative.

Recent research investigates on deep learning technologies for automated analysis of people in video sequences. The basic technology of this domain is the Convolutional Neural Network (CNN) which is used in many different computer vision tasks as a supervised learning algorithm. This algorithm can be used to detect visual features of given images and calculates an output vector which corresponds to the Big-Five personality model.

The "First Impressions Challenge 2016", organized by ChaLearn Looking at People (LAP) [3] was held in order to promote research on the automatic determination of personalities in videos. The results of the winner teams [20][16][7] have been published at the European Conference on Computer Vision 2016 (ECCV 2016) at the Workshop *Challenge on Apparent Personality Analysis*.

The results show that the automatic assignment to the Big-Five personality model using CNN based systems achieves promising results by using a multi-modal system consisting of visual-based and audio-based information. The used methods of [20][16][7] differ significantly in terms of system architecture, model training techniques and data pre-processing. It is not clear to which extent, different aspects affect the performance of the introduced systems.

The purpose of this work is to investigate this matter by starting from a CNN-based baseline solution and extending this solution in different directions. The main aim of this paper is to show how different pre-processing methods, like face-extraction and data-augmentation[13], influence the predicted personality confidences. Furthermore, it explores different deep learning technologies e.g. reg-

*daniel.helm@tuwien.ac.at
†martin.kampel@tuwien.ac.at

ularization in order to improve the performance of the proposed models. Moreover, the influence of different content-based visual information, like face-features and extended-image-features are explored. Finally, the paper points out the complexity of developing and training a CNN based system to solve a regression task.
The paper is structured in nine sections. In Section 1, the motivation and the problem description are introduced. Similar problems and alternative approaches are discussed in Section 2. In section 3 a detailed overview of the used pre-processing techniques is given. The implementation details and the used evaluation metrics are discussed in section 4 and 5. The subsequent section 6 presents and analyzes the *Basis Convolutional Neural Network*. An alternative approach is discussed in the section 7 *3D Convolutional Neural Network*. Finally, section 8 points out the conclusion of the results and presents future work. The paper closes with the acknowledgment in the last Section 9.

2 Related work

In the last decade, many different approaches and models for the automated evaluation of personality traits have been published. Therefore, different information channels like audio-based [17][1], text-based [2][1] or visual-based [14][6] features are used to find correlations to personality models. [15][5] publish approaches of combining different information channels, so-called multi-modal systems, to get the models more robust.

The winning teams of ChaLearn LAP [20][7][16] propose a multi-modal system which combines audio as well as visual features to predict the Big-Five personality traits. The proposed system of the first winning team [20] is based on a VGG-Face model [12] and is extended with an Extended Descriptor Aggregation Network (DAN+)[19]. The extracted audio features are related to the personality model by using log-filter-bank analysis [4] and a neural network[18]. Data fusion of the calculated Big-Five confidences is done by simply building the mean average values. The approach of the second winner team [16] is based on a combination of a 3D-CNN [10], a Long-Short-Term Memory (LSTM)[9] and a neural network[18]. [7] presents a multi-modal system based on a Deep Residual Network (ResNet)[8].

3 Pre-Processing

3.1 Dataset and labels

The dataset used in this work is provided by ChaLearn Looking at People (LAP)[1]. It consists of 10000 videos which are separated into three parts. The training set includes 6000 videos, the validation, and test set both include 2000 videos. Each video shows a person telling something about him or herself. The duration of each video is about 15 seconds with a frame rate of about 30 fps. The training set also includes the ground truth labels which represent the Big-Five confidences for each video. The ground truth labels were obtained by workers of Amazon Mechanical Turk (AMT). Detailed information is mentioned in the paper [3]. The organization ChaLearn LAP didn't publish the labels for the test and validation set during the evaluation phase of the described algorithms in this paper. For this reason, the 6000 videos of the training set were used for developing as well as evaluating the created system by splitting it to 4800 training and 1200 validation videos. Figure 1 shows three randomly selected examples of the dataset.

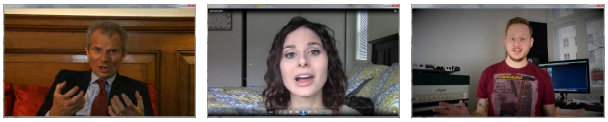


Figure 1: Three randomly selected examples of the ChaLearn LAP dataset

3.2 Face-feature vs. Extended-Image-Feature extraction

As the input data of the system are videos, it is necessary to select image frames of each video first in order to extract content information. The frame rate of the original videos is about 30 frames per second and was reduced to 4 fps. This is done due to resource constraints and enabling image extraction in continuous time distances without losing temporal video information. For each video, there are about 60 images available, whereas each image shows the person as well as the background. Since the goal is to predict the personality traits, the information of the background is not needed and can have a negative influence on the results. Because of this reason, just the face in each image is extracted. Figure 2 shows the pipeline of the face extraction process. However, also body gestures like hand movements or the clothes can have a significant impact on the first impression of a person. Therefore, extended image features are extracted and evaluated as well. Figure 3 shows the calculation process of the extended region of interest with a sample image of the dataset.

The open source library Dlib[2] is used to detect a 68-dimensional feature vector which describes the facial landmarks. With this information, it is possible to rotate the face so that the intraocular distance of the eyes is horizontal to the x-axis of the image. After rotating, each face is cropped by creating a bounding box with the center-point

[1] http://chalearnlap.cvc.uab.es/dataset/20/description/ last visit: 01/22/19

[2] http://dlib.net/ last visit: 02/05/19

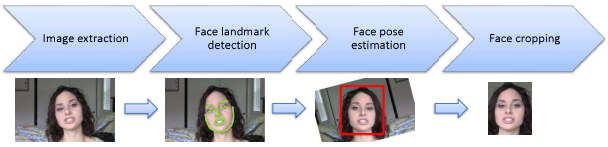


Figure 2: The pre-processing of each video is done with the proposed *Face extraction pipeline* in order to achieve a dataset where the faces of each extracted image are in the same position and pose.

of the facial nose. Finally, the cropped face is resized to 64x64x3. The ambient light of the extracted face images strongly varies, therefore histogram equalization is used to increase the global contrast of each image.

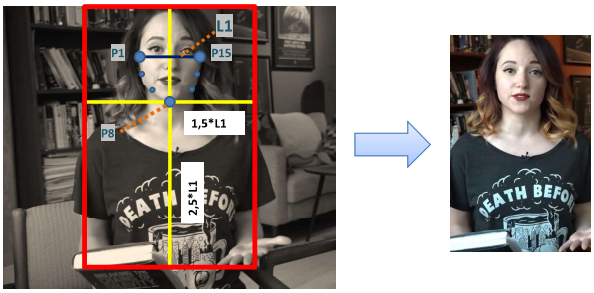


Figure 3: The left image shows one extracted frame of a video in the dataset and illustrates the calculation of the bounding box. The right image displays the resulting crop of this example.

3.3 Mean substraction and normalization

Mean substraction (zero-centering) and normalization [3] are significant processes in the context of computer vision and deep learning. Mean substraction is needed to transform the individual lighting conditions of the images in the training set into one global setting without influencing the content of the images. Therefore, the mean value for each color channel is calculated of the training set. In the next step these mean values are substracted of all images in the training, validation and test set. Afterward, the standard deviation of the zero-centered training set is calculated. In the last step the normalized images of the datasets are divided by the calculated standard deviation of the training set. Finally, this process results in normalizing the contrast to have a standard distribution with a variance equal to one. Figure 4 points out the effect of mean substraction and normalization.

3.4 Data augmentation

The training of a Convolutional Neural Network depends significantly on the dataset size. Since calculating Big-

[3] http://cs231n.github.io/neural-networks-2/ last visit: 02/06/19

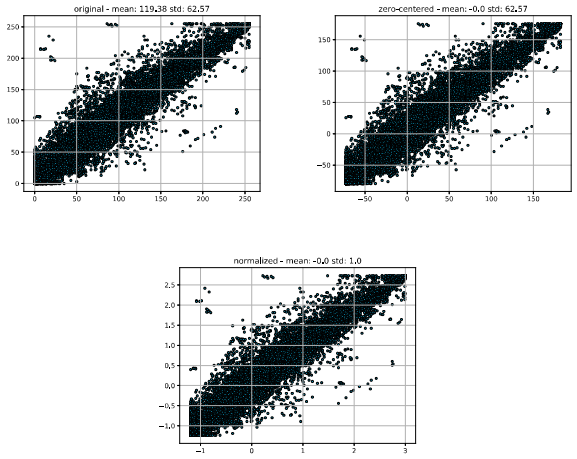


Figure 4: The *upper left graph* points out the distribution of the original dataset. The *upper right graph* displays the zero-centered data by subtracting the mean value. The normalized data in the *lower centered graph* are achieved by dividing the standard deviation. The graphs show only the distribution for the red color channel.

Five confidences is a regression problem which is challenging to handle, more images help to get a better training effect of the model. For this reason, data augmentation is used to produce additional images on-the-fly in order to increase the model performance. In this paper three methods for data augmentation are used: Horizontal flipping, random rotations and random crops (Figure 5).

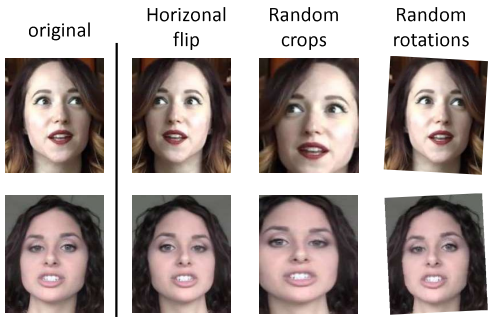


Figure 5: This figure displays two examples of the used *Data augmentation mechanisms* in this paper.

4 Implementation details

The proposed algorithms are implemented using the framework Keras[4], a Python library which uses Tensorflow[5] as backend. The model training takes about two days for 200 epochs using the operating system Linux Ubuntu

[4] https://keras.io/ last visit: 02/09/19
[5] https://www.tensorflow.org/ last visit: 01/29/19

16.04 LTS[6] and one GPU Palit Geforce GTX 1060[7].

5 Evaluation

The evaluation of the models is done by using the given metrics of the organization ChaLearn LAP during the "First Impressions Challenge 2016". Therefore, the average accuracy of each individual personality trait (1), as well as the mean average accuracy (2), is calculated to get a single representation of the model performance. gt_{ij} represents the ground truth predictions for each video whereas p_{ij} stands for the calculated predictions. N is the number of available videos and m counts the dimensions of the Big-Five personality model. This procedure is necessary to compare the results derived in this paper with the results of the winning teams of the challenge. A detailed description of the evaluation metrics can be found in [3].

$$AverageAccuracy_j = \frac{1}{N}\sum_{i=1}^{N}(1 - \left|gt_{ij} - p_{ij}\right|) \qquad (1)$$

$$MeanAverageAccuracy = \frac{1}{m}\sum_{j=1}^{m}(AverageAccuracy_j) \qquad (2)$$

6 Basis Convolutional Neural Network

6.1 Model architecture

The Basis Convolutional Neural Network (CNN) consists of two general parts, the feature extraction, and the regression part. This architecture is based on the VGG-16 model [12] which is conceived to resolve a classification task. Thus, the model is optimized to solve a regression problem by using the sigmoid activation function at the last fully connected layer (fc2) and the loss function Mean Squared Error (MSE) instead of the Crossentropy function. The activation function *Rectified-Linear-Unit (ReLu)* as well as *Batch-Normalization* are applied to all convolutional layers and fc-layers except the output-layer. A batchsize of 64 is selected. Furthermore, the Basis-CNN model is only able to predict image-based Big-Five confidences. Therefore, post-processing is implemented by calculating the mean values of all predicted Big-Five confidences corresponding to one video. Figure 6 shows a schematic setup of network architecture with about 6.5 million model parameters.

The training of the model is done with the Stochastic Gradient Descent (SGD) method as well as Binary

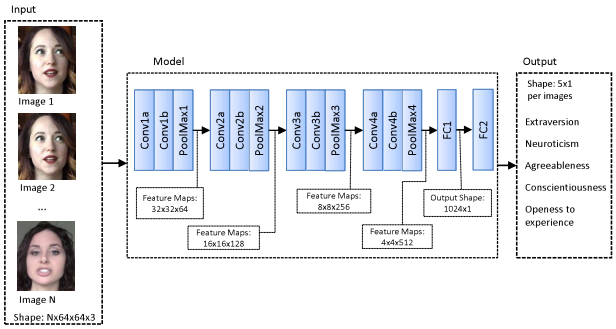


Figure 6: The system architecture of the proposed Basis Convolutional Neural Network

Crossentropy function. The model is trained several times with different hyperparameters, and with and without data augmentation. Each training runs 200 epochs and the dataset is shuffled by each epoch to get better regularization effect of the model.

6.2 Results

Four defined experiments 1,2,3 and 4 are compared and evaluated. Experiment 1 and 2 explore the effect of using and not using data-augmentation during the training phase. Dropout probability and L2-distance regularization (weight decay) are special techniques to improve model performance and avoid overfitting. Experiment 3 and 4 point out how these parameters influence the model performance in the context of a regression problem.

Table 1 gives an overview of the defined experiments. Furthermore, the final Big-Five predictions of the extracted face-features are evaluated by using the loss function Binary Crossentropy and MSE. Additionally, in experiments 2, 3 and 4 the learning rate is reduced by a factor of 10 if the validation loss is not decreasing in the last 30 epochs.[8]

Experiment	Description
1	without data-augmentation
2	with data-augmentation
3	+ Weight Decay
4	+ Dropout

Table 1: An overview of defined experiments explored in this paper.

Figure 7 shows the training and validation history of the Basis-CNN model over 200 epochs for each experiment.

Experiment 1 shows that the training loss values 7a decrease with each epoch during the training phase. The corresponding validation loss curve 7b points out that the values begin to flatten after first few epochs and with the beginning of epoch 21 an increase of the validation loss values can be monitored. One mechanism to counteract

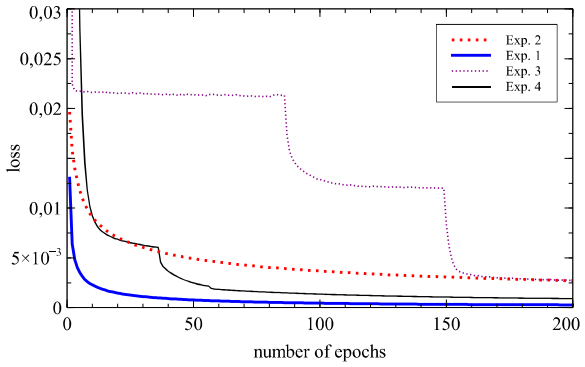
[6] http://releases.ubuntu.com/16.04/ last visit: 12/15/18

[7] http://www.palit.com/palit/vgapro.php?id=2674&lang=en last visit: 02/03/19

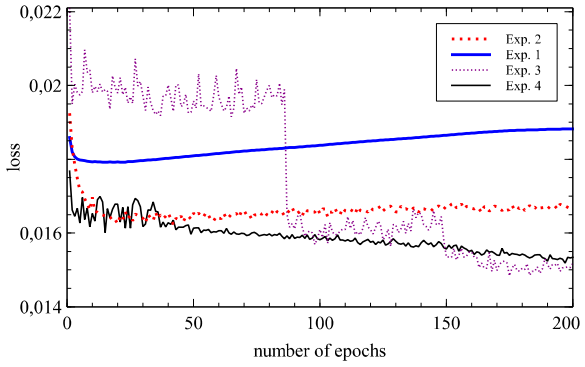
[8] http://cs231n.github.io/neural-networks-2/ last visit: 02/06/19

the over-fitting effect is to use data-augmentation. The resulting loss history of experiment 2 reveals the effect by using this mechanism which shows the first improvements of the model performance. The improvement is due to the increased data size elicits throw the data-augmentation technique.

Experiment 3 is an enhancement of experiment 2 and uses L2-distance regularization (global weight decay: 0,002) and learning rate reduction technique (initial-value: 0,1). The learning rate of Stochastic Gradient Descent is reduced by a factor of 10 if the validation loss values do not change significantly over a duration of 15 epochs. The effect on the loss history of experiment 3 can be seen at epoch 80 and 150, which shows a continuous training effect of the model during the whole training phase. A further hyper-parameter is introduced in experiment 4 called Dropout probability (value: 0.2). This mechanism affects the fully connected layers of the regression part of the model. Therefore, connections between the single neurons of each layer are dropped in order to the selected dropout probability. Experiment 4 reveals similar accuracies to experiment 3 without a significant performance increase. Thus, those two experiments show the most promising results of the evaluation by using MSE (see table 2).



(a) Training Loss



(b) Validation Loss

Figure 7: An overview of the training and validation loss history of experiment 1, 2, 3 and 4 trained with Mean Squared Error.

Based on the observed results in the previous experiments a further experiment is set up. It evaluates the effect of using extended-images-features on model performance instead of face-features. Table 3 shows the final accuracies of the model trained with face- and extended-image features. The most promising results are achieved by using Binary Crossentropy loss function in combination with sigmoid activation at the output layer and training the model with the extended-image-features.

Exp.	E	A	C	N	O	Mean acc.
Bin.Cross.	0,8841	0,8947	0,8786	0,8801	0,8881	**0,8851**
MSE	0,8786	0,8939	0,8751	0,8779	0,8836	**0,8818**
[20]	0,9133	0,9126	0,9166	0,9100	0,9123	0,9130
[16]	0,9150	0,9119	0,9119	0,9099	0,9117	0,9121
[7]	0,9107	0,9102	0,9138	0,9089	0,9111	0,9109

Table 2: An overview of the achieved test dataset results compared to the three winner teams of the "First impressions Challenge 2016".

Exp.	E	A	C	N	O	Mean acc.
Ext. Feat.	0,8857	0,8951	0,8815	0,8829	0,8894	**0,8869**
Face Feat.	0,8841	0,8946	0,8786	0,8801	0,8881	**0,8851**

Table 3: The test set results of the proposed models trained with the extended-image-features compared to the face-features.

Finally, a model accuracy of 0,8869 is reached with the extended-image-features. However, the proposed results don't reveal significant improvements. For a better understanding of the results, feature maps of randomly selected frames are extracted from the last *MaxPooling layer* of the network. To compare this feature maps directly with the original input frame *unpooling* and *deconvolution* are applied to get the same dimensions. The feature-map analysis of the proposed CNN model points out that the model tends to weight only feature areas on the left side with higher priority which are not always relevant for the first impression of a person. Figure 8 compares extracted feature maps related to extended image features and face-features. It can be concluded that the defined model architecture and the selected training strategies are not an optimal solution to solve the proposed problem in this paper.

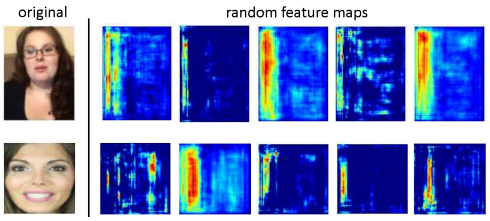


Figure 8: This figure points out extracted feature maps of two example images of the extended-image-features and the face-features.

7 3D Convolutional Neural Network

7.1 Model architecture

The 3D Convolutional Neural Network (3D-CNN) is able to calculate video-based predictions of the Big-Five personality model. Unlike the Basis-CNN model, no separate data fusion has to be performed. One more advantage is the ability of the model to consider the temporal aspects of the given videos. Figure 9 shows the proposed model architecture.

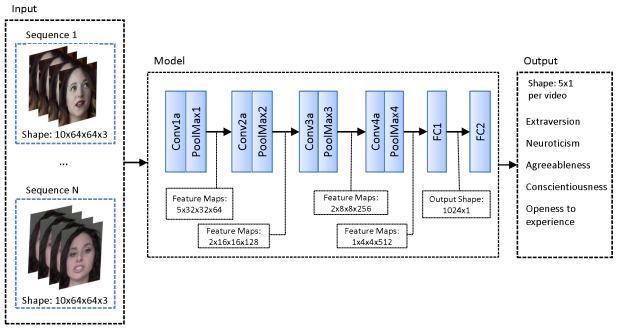
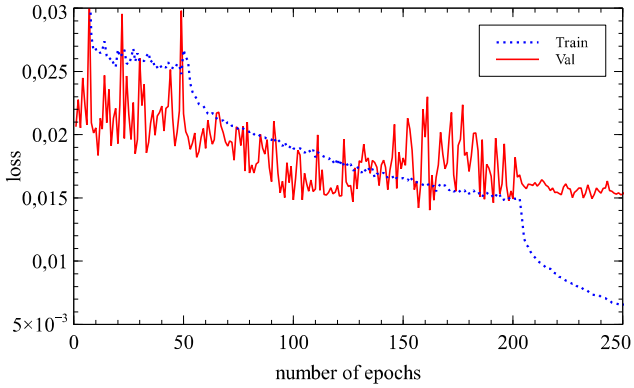


Figure 9: The system architecture of the proposed 3D Convolutional Neural Network model.

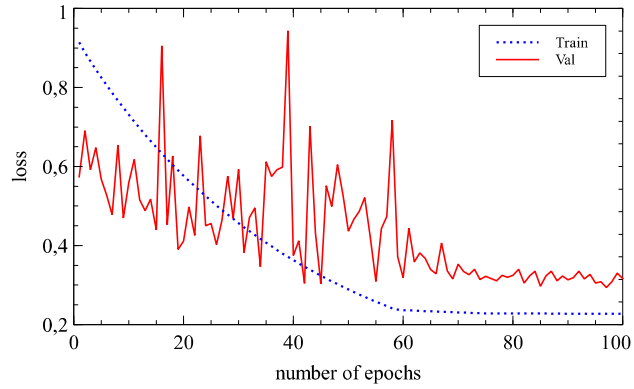
Two different training strategies are used to train the 3D-CNN model. Training strategy 1 explores the effect of the Binary Crossentropy function and training strategy 2 evaluates the Mean Squared Error function. Both strategies are trained with a maximum of 300 training cycles and evaluated with different hyperparameter selections in an empirical way. The model consists of about 12,5 million parameters. Furthermore, the mechanism *early stopping* is used to stop the training process if the performance on the validation set is not increasing for 20 epochs. The learning rate reduction, global weight decay and dropout probability are equal to the training strategy of the Basis-CNN model. Furthermore, the extracted image-sequences consisting of 60 frames per video are split into 10 non-overlapping temporal parts. Finally, one frame of each temporal part is dynamically extracted after each training cycle based on the idea of [16].

7.2 Results

Figure 10 gives an overview of the training loss history of training strategy 1 and 2. The results show significant differences in training as well as validation loss during the whole training period compared to the proposed strategies for the Basis-CNN model. During each training cycle, the 3D-CNN model extracts new images for each video. Thus, the set of input images varies which is the main reason for the fluctuating validation loss history. Because of these observations it can be assumed that the proposed training strategy yields other challenges.



(a) Training strategy 2 - MSE



(b) Training strategy 1- Binary Crossentropy

Figure 10: An overview of the training and validation loss history of training strategy 1 compared to strategy 2.

Table 4 shows the resulting accuracies by evaluating the final test dataset. The results display a slight increase of the performance using the Mean Squared Error loss function trained with face-features but point out no significant difference.

Exp.	E	A	C	N	O	Mean acc.
Strat.1 - BC	0,8546	0,8771	0,8573	0,8545	0,8600	0,8607
Strat.2 - MSE	0,8640	0,8827	0,8668	0,8655	0,8726	0,8703

Table 4: The results of training strategy 1 compared to strategy 2 evaluated with the final test set.

As well as the Basis-CNN model, the 3D-CNN model is trained with the extended-image-features. Thus, it is possible to compare the achieved results to those of the Basis-CNN model. The accuracy of the 3D-CNN model trained with extended-image features reaches a value of **0.8905** and shows the most promising results of the published approaches in this paper. It has been found that the 3D-CNN model is more robust than the Basis-CNN model by evaluating the distribution of the calculated Big-Five confidences of the test dataset. Figure 11 compares the distribution of the calculated Big-Five confidences with the Basis-CNN model and the winning team of the ”First

Impressions Challenge 2016" [20] as well as with the distribution of the ground truth data. Moreover, it can be argued that the 3D-CNN model, trained on the extended-image-features, learned features of more relevant areas of the images. Table 5 summarizes the achieved results of the 3D-CNN model trained with face-features and extended-image-features.

Exp.	E	A	C	N	O	Mean acc.
Face-Feat.	0,8771	0,8870	0,8699	0,8727	0,8817	0,8777
Ext.-Feat.	0,8883	0,8986	0,8886	0,8856	0,8916	**0,8905**

Table 5: The results of the proposed 3D-CNN trained with the face-features compared to the extended-image-features.

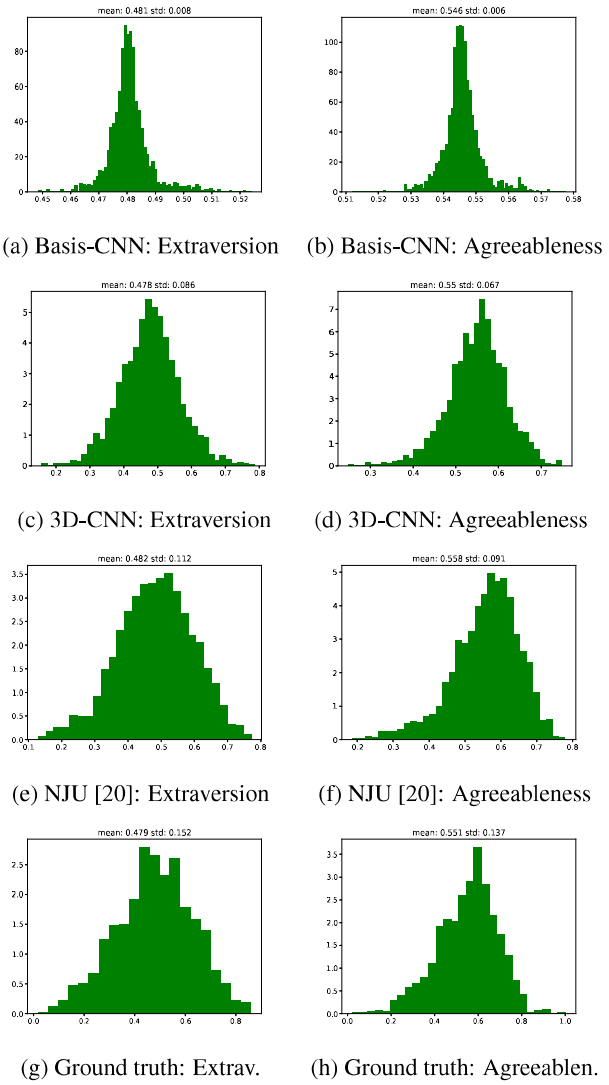
8 Conclusions

In this paper, two models are presented which are able to predict the Big-Five personality model by using short input video sequences. The proposed models are based on the VGG model architecture and show promising results. With the displayed algorithms in this paper accuracies up to 0,8905 are observed by using visual image features and no audio features such as the winning teams [20][16][7] of the "First Impressions Challenge 2016" [3].

A qualitative analysis of the proposed Basis-CNN points out that the evaluated techniques, like Dropout probability or L2-Distance regularization, do not show significant performance improvements. Moreover, the evaluation of the standard deviation in the Basis-CNN is significantly smaller than in the proposed 3D-CNN model. Finally, a detailed analysis of the feature maps of the Basis-CNN model displays that the network architecture learns features only from less relevant areas in the input images. Therefore, it can be concluded that the proposed Basis-CNN architecture should be observed very critical in the context of this paper.

The explored 3D-CNN architecture achieves the best overall accuracy of 0,8905 compared to all other experiments in this paper. Furthermore, the evaluation of the Big-Five confidence distribution shows a more reliable standard deviation compared to the winning teams of the "First Impressions Challenge 2016"[20][16][7]. In the context of predicting Big-Five personality confidences, it can be argued that the proposed 3D-CNN architecture is the most promising algorithm compared to the other explored models in this paper. This architecture can be enhanced by exploring the audio features as well. Furthermore, improvements, such as optimization of the model architecture and training strategies, are planned for future work.

Finally, the proposed methods are a very basic starting point in order to develop an automatic video analysis system such as a rhetorical or personality trainer. Furthermore, the ground truth labels of the used dataset have to be explored very critical because they are based on the estimation of a limited number of experts. Moreover, all kinds of systems or methods which are able to predict the first impression or the personality of human beings raises ethical questions. They could be used to discriminate persons in different situations such as job applications, sales hearings or immigration.



(a) Basis-CNN: Extraversion (b) Basis-CNN: Agreeableness

(c) 3D-CNN: Extraversion (d) 3D-CNN: Agreeableness

(e) NJU [20]: Extraversion (f) NJU [20]: Agreeableness

(g) Ground truth: Extrav. (h) Ground truth: Agreeablen.

Figure 11: This figure displays the calculated distributions of the predicted Big-Five confidences by applying the final test dataset.

9 Acknowledgement

I would like to thank my supervisor Martin Kampel and co-supervisor Christopher Pramerdorfer for the good cooperation, the great support, and their patience. Furthermore, a special thanks I want to express to Lisa Glatzer for all the valuable feedback she offered during the creation of this work.

References

[1] F. Alam and G. Riccardi. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 955–959. IEEE, May 2014.

[2] Firoj Alam, Evgeny A Stepanov, and Giuseppe Riccardi. Personality traits recognition on social network-facebook. *WCPR (ICWSM-13), Cambridge, MA, USA*, pages 1–4, 2013.

[3] Escalera Sergio Clap Albert, Escalante Hugo Jair. Chalearn lap 2016 : First round challenge on first impressions - dataset and results. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops: Proceedings, Part III*, pages 400–418. Springer International Publishing, 2016.

[4] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, Aug 1980.

[5] Golnoosh Farnadi, Shanu Sushmita, Geetha Sitaraman, Nhat Ton, Martine De Cock, and Sergio Davalos. A multivariate regression approach to personality impression recognition of vloggers. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition*, WCPR '14, pages 1–6. ACM, 2014.

[6] TGI Fernando et al. Persons personality traits recognition using machine learning algorithms and image processing techniques. *Advances in Computer Science: an International Journal*, 5(1):40–44, 2016.

[7] Yamur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven, and Rob van Lier. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops: Proceedings, Part III*, pages 349–358. Springer International Publishing, 2016.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, June 2016.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[10] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[11] Robert R. McCrae and Paul T. Costa. A contemplated revision of the NEO Five-Factor Inventory. *Personality and Individual Differences*, 36(3):587–596, 2004.

[12] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In Xie Xianghua, Mark W. Jones, and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 41.1–41.12. BMVA Press, September 2015.

[13] Christopher Pramerdorfer and Martin Kampel. Facial expression recognition using convolutional neural networks: State of the art. *arXiv preprint arXiv:1612.02903*, pages 1–6, 2016.

[14] Rizhen Qin, Wei Gao, Huarong Xu, and Zhanyi Hu. Modern physiognomy: An investigation on predicting personality traits and intelligence from the human face. *arXiv preprint arXiv:1604.07499*, pages 1–27, 2016.

[15] Maxim Sidorov, Stefan Ultes, and Alexander Schmitt. Automatic recognition of personality traits: A multimodal approach. In *Proceedings of the 2014 Workshop on Mapping Personality Traits Challenge and Workshop*, MAPTRAITS '14, pages 11–15. ACM, 2014.

[16] Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops: Proceedings, Part III*, pages 337–348. Springer International Publishing, 2016.

[17] Fabio Valente, Samuel Kim, and Petr Motlicek. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *Thirteenth Annual Conference of the International Speech Communication Association*, pages 1183–1186. ISCA, 2012.

[18] Sun-Chong Wang. *Artificial Neural Network*, pages 81–100. Springer US, Boston, MA, 2003.

[19] X. S. Wei, J. H. Luo, J. Wu, and Z. H. Zhou. Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing*, 26(6):2868–2881, June 2017.

[20] C Zhang, Hao Zhang, X Wei, and Jianxin Wu. Deep bimodal regression for apparent personality analysis. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops: Proceedings, Part III*, pages 311–324. Springer International Publishing, 2016.