First impression based personality analysis

Jelena Gorbova

Project final report
Neural Networks course (LTAT.02.001)

1 Introduction

In the past few years human behavior has became a topic of high interest in computer vision field. Many researchers are still focusing on the problem of how to teach computers to identify people by face, detect their gestures, facial expressions or recognize their emotions. Personality automatic analysis was less observed until the recent, even as it could find applications in many different areas, such as security and candidate selection problems.

Personality affects first impression which person leaves by communication with other people, which in its turn affects decisions people make, for example by deciding whether we like or dislike person, or choosing the right candidate for a job, since in this case personality characteristics play a role on equal basis with candidate professional skills. For testing professional skills participants can be given test assignments, but it's very time-consuming to interview each candidate in person. The algorithm, which would provide the relevant information about personality characteristics of each candidate could save a lot of time and human resources for solving the above mentioned problem.

Automatic personality analysis has gotten more attention in computer vision field under challenges organized by ChaLearn Looking at People (Chalearn LAP) group (*1*). In 2014 ChaLearn LAP has published a First Impression dataset, which contains short video clips with corresponding

5 Big Personality Traits scores.

The First Impression database contains 10000 video-clips taken from more than 3000 different HD YouTube videos, where people are mostly sitting and speaking in English in front of the camera in very different lighting conditions and background scenes. People in the videos belong to different age, gender, nationality and ethnic groups. Moreover the database represents some exceptional cases, e.g. on some videos people were speaking sign language. There are also cases when person is sitting in front of the camera without movement and uttering a singe word.

Each video is labeled with 6 values in the range from 0 to 1. Five of them describe the 5 big personality traits, namely extroversion, agreeableness, conscientiousness, neuroticism and openness. With the 6^{th} value (Interview) Amazon Mechanical Turk (AMT) workers estimated whether the person in the video clip should be invited to a job interview or not. In this project only 5 first values were used, since the interpretation of Interview score is very doubtful and is out of projects field of interest.

In 2016 two challenge rounds had been held using this dataset, in which participants developed solutions for recognizing personality traits of users in short video sequences (*2, 3*). In 2017 Chalearn LAP has organized the additional challenge round where they have brought up job candidate selection problem (*4*). Under the last round participants were asked to predict the score of job interview invitation beside the personality scores.

The 5 Big Personality Traits used in the above mentioned dataset are widely used in psychology for characterizing the major personality properties of a human being. These can be listed as follows:

- **Extroversion** (sociability, assertiveness);
- **Agreeableness** to other people (friendliness);
- **Conscientiousness** (discipline);

- **Neuroticism** (emotional stability);
- **Openness** to experience (intellect).

In this paper author presents a system of first impression based automatic personality screening from short video presentations by using visual modality. Drawing on the previous studies it's taken into account, that it's possible to implement the personality analysis without personal contact. Based on that here is presented a system aimed to estimate a persons scores in above mentioned personality characteristics using combination of convolutional neural network (CNN) for image features extraction and recurrent neural network (RNN) architecture for learning of temporal pattern.

In Section 2 the personality automatic analysis related works are presented. In Section 3 a full description of proposed method is provided. In Section 4 experimental results is presented.

2 Related works

As it was already mentioned in previous section the automatic personality analysis has became a very relevant topic in field of computer vision in the past few years. Some approaches proposed by challenge (*2–4*) participants are presented in this section.

Wei at al. have participated in the Fisrt Impression challenge in 2016 year (*2*), where they achieved the accuracy over 0.91 for all 5 traits on the test set. In their work (*5*) they propose a bimodal approach for prediction of 5 Big Personality Traits scores based on visual and audio input. For visual modality they have used the modified VGG-face architecture. They discard the fully connected layers,replaced them by both average- and max-pooling following the last convolutional layers. Each pooling operation is followed by the standard l2-normalization. For audio modality they extract Mel Frequency Cepstral Coefficients (MFCC) and logfbank features, which are fed to model composed of a fully-connected layer followed by a sigmoid function layer

to train the audio regressor. The final accuracy is obtained by taking the average of both modalities predictions.

In (*6*) authors propose two architectures personality automatic analysis. Same as Wei at al. they fuse audio and visual features to learn the temporal information. The first methodology proposed in (*6*) uses Volumetric (3D) convolution based deep neural network, while the second one is formulated with an LSTM (Long Short Term Memory) based deep neural network. Both approaches have very deep and complex architecture and include convolutional image data processing. Based on test set accuracy the LSTM based approach achieves in general better results. The averaged accuracy for LSTM and Volumetric based networks are 0.913 and 0.912 respectively.

In (*7*) Grpinar et al. present a multimodal approach, which includes not only face area and audio data processing, but also uses the video background (scene) information. For facial features extraction they fine-tune the VGG-face model changing the final layer to a 7-dimensional emotion recognition layer using more than 30K training images in the FER-2013 dataset. For scene features extraction they use another pretrained model, namely VGG-VD-19 network, which was trained for an object recognition task on the ILSVRC 2012 dataset. Authors of proposed method have participated in the second round of First Impression challenge (*3*) and they have achieved the accuracy over 0.912 for all 5 personality traits.

3 Proposed method

At this stage several approaches were developed to process a numerical conclusion from video input (*8*). One of the most commonly used is a combination of convolutional features and RNN. This approach was implemented in this project to predict 5 personality scores based on short video input. Generally the proposed method consists of 2 parts:

- Preprocessing

- Feature extraction and learning

A block diagram of the proposed method is shown in Figure 1.

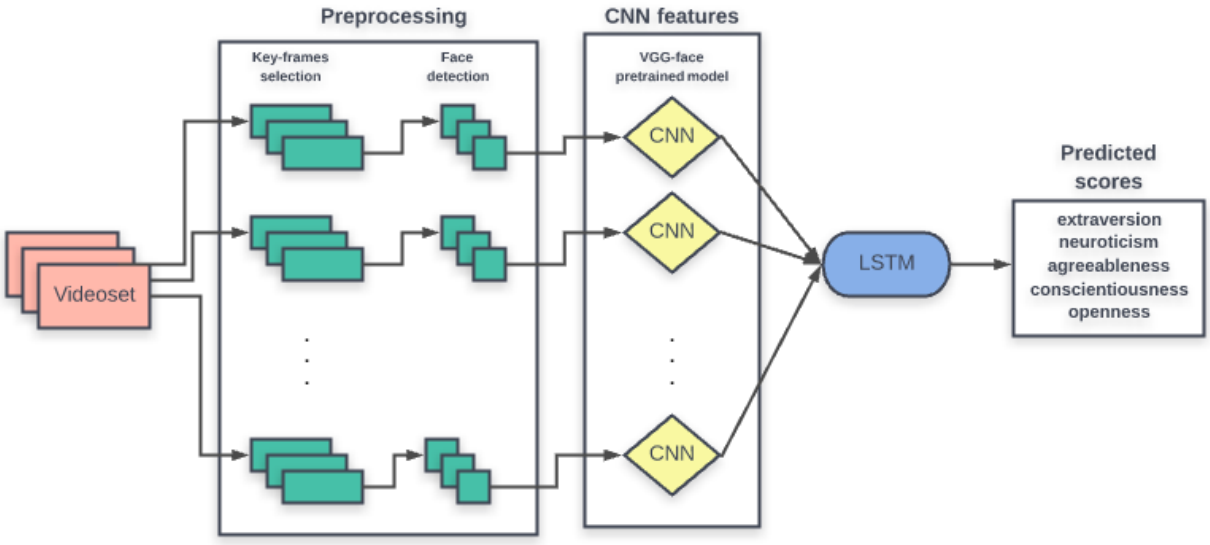

Figure 1: The block diagram of the proposed method. Row video-input is red, extracted key-frames are green

3.1 Preprocessing

Preprocessing part carries two main aims in this project: i) Reduce input dimensionality and memory cost; ii) make the input more informative. The dataset, which was used in this work, contains videos with different length and sample-rate. 15 key frames were selected to represent each video by fixed size set of frames and reduce the input dimensionality. Firstly each video was divided into 15 non-overlapping segments. After that in each segment were found a frame, which pixel-wise was the nearest to the segments centroid. Let us observe the certain input video and denote video i-th segment by S_i. The centroid of this segment is calculated as averaged pixels of all frames in S_i:

$$centroid_j = \frac{\sum S_{ij}}{N_i},$$

where j defines a color channel, N_i total number of frames in S_i. The key frame of S_i is found as the closest to $centroid$ frame based on euclidean distance, which is calculated by following formula:

$$d(frame_i, centroid) = \sqrt{\sum (frame - centroid)^2}.$$

At the second stage the face region was registered from each key frame. Many libraries such as opencv (*9*) or dlib (*10*) propose their built functions for face region(s) detection. The opencv face detection algorithm uses Haar Feature-based Cascade Classifiers, while dlib is based on Histogram of Oriented Gradients (HOG) features combined with a linear classifier. First Impression database videos have different qualitative characteristics and opencv algorithm is very sensitive to the arguments changes (e.g. number of nearest neighbors and image scale), hence dlib face detection function was used in this project.

3.2 CNN features extraction

Convolutinal neural networks (CNN) are widely used for emotion and face recognition. One of the most well-known models VGG-Face was presented by Parkhi at al. (*11*). Originally it was aimed to recognize faces from image input. It is trained on 2600 individuals with around 3 million images and has very deep and complex architecture (see Figure 2). In (*11*) is reported about achieving the accuracy over 97% tested on Youtube Faces Dataset.

Under this project there was made an assumption, that features which are successfully used for face recognition tasks can provide relevant information for personality analysis. The VGG-Face (*12*) pretrained model was used in this work for high-level feature extraction. Empirically fc7 features (4096 dimensional) were chosen to represent each key-frame on feature level.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	–	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	–	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	–	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256
num filts	–	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512
stride	–	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	–	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmx
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	prob
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	–	512	–	512	–	–	512	–	512	–	512	–	–	512	–	4096	–	4096	–
num filts	–	512	–	512	–	–	512	–	512	–	512	–	–	4096	–	4096	–	2622	–
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

Figure 2: Architecture of VGG-Face CNN model

3.3 LSTM

The long short time memory (LSTM) is a type of recurrent neural networks first presented in (*13*), the method also takes into account long-term dependencies, instead of just short-term ones. The LSTM cell takes 3 types of inputs that are guarded by the forget (f), input (i) and output (o) gates.

Let denote the hidden output of time moment t by h_t, then input by x_t, the cell state by c_t and all the gates similarly f_t, i_t, o_t. The initial values of c_0 and h_0 are set to 0. Each gate $g \in \{f, i, o\}$ is characterized by matrix W_g, U_g and b_g. Similarly the cell is described by W_c, U_c, b_c. Here W denotes weight for input x_t, U carries weights for the hidden output of previous time moment and b is the bias term of corresponding gate or cell.

This way the gates for each time step are characterized by:

$$g_t = S(W_g x_t + U_g h_{t-1} + b_g),\ g \in \{f, i, o\}$$

The corresponding cell state c_t is calculated as:

$$c_t = f_t * c_{t-1} + i_t * \textit{tanh}(W_c x_t + U_c h_{t-1} + b_c)$$

Here $*$ denotes the Hadamard alos known as the element-wise product of two matrices. The output for the cell is found with:

$$h_t = o_t * \textit{tanh}(c_t)$$

In this work many-to-one LSTM architecture was used to learn the temporal information from CNN features sequences and predict 5 personality scores (see implementation details in Section 4).

4 Experimental results

Firstly the "First Impression" dataset were splitted to train (6000), validation (2000) and test (2000) set in the same way as in ChaLearn Lap CVPR/IJCNN Competition Challenge 2017. That allows us to compare results with challenge participants. Testing and selecting the best model was done using the training and validation set.

As it already was mentioned the face detection were implemented using dlib built function and was for 99% automatic. In some exceptional cases dlib wasn't able to detect face automatically and it was done manually.

Feature extraction and temporal learning were implemented using Tensorflow library. The pretrained model VGG-face was used for feature extraction. Empirically based on validation set performance was chosen the model with two recurrent layers and 10 hidden nodes. The learning rate was set to 0.0001 and batchsize to 1000. Adam optimization method was used during the training and mean absolute error (MAE) as cost function.

The final performance rates in each category are presented in Table 1, these were calculated with the following formula:

$$accuracy = 1 - MAE = \frac{\sum_{i=1}^{N_t} (1 - |p_i - r_i|)}{N_t}$$

where N_t is the number of videos in the test set and p_i and r_i are the predicted and real values, respectively. The same accuracy metric was used in the ChaLearn Lap CVPR/IJCNN Competition Challenge 2017.

As can be seen in Table 1 proposed method results in less accuracy precision than top three

Label	**Proposed method**	heysky (I)	Bekhouche (II)	go2chayan (III)	azzasama
Extroversion	**0.8809**	0.9213	0.9155	0.9027	0.8788
Neuroticism	**0.8783**	0.9146	0.9083	0.9011	0.8632
Agreeableness	**0.8952**	0.9112	0.9103	0.9032	0.8721
Conscientiousness	**0.8733**	0.9152	0.9138	0.8949	0.866
Openness	**0.8858**	0.9170	0.9101	0.9047	0.8748

Table 1: Comparison of prediction accuracy on test set for 5 personality traits with ChaLearn Lap CVPR/IJCNN Competition Challenge 2017 participants

challenge participants. Only for agreeableness the accuracy of prediction is over 0.89. The averaged accuracy for all 5 personality traits scores is 0.882, hence the proposed method results into higher accuracy only in compare with the last place in the mentioned challenge.

The relatively low prediction accuracy can be caused by fact, that top three challenge participants have used much more complex approaches. The first and third places used multiple modalities, e.g. audio, background features or lexical context, which implementation requires more technical, time and human resources.

5 Discussion and Conclusion

In this project video processing system was presented for the first impression personality analysis. The pretrained CNN model VGG-face was used for high-level convolutional features extraction. After that extracted features were fed to LSTM network to predict final scores. Comparing obtained results with ChaLearn Lap CVPR/IJCNN Competition Challenge 2017 participants the proposed method results to lower efficiency. But considering, that proposed method uses very simple tools and convolutional features extraction were done with a model, which was trained for face recognition tasks, the result presented in this work can be considered as reasonably accurate.

Presented system surely can be improved fine-tuning VGG-face pretrained model, since it was originally used for face recognition tasks. Assuming that emotion recognition is more similar to personality analysis problem one of the suggested improvements is model fine-tuning with any

emotions database. Also in this project each video was represented by 15 key frames (ca 1 frame per second). Possible, that increasing the number of key-frames can provide additional information and improve algorithm performance.

References

1. "Chalearn looking at people." [Online]. Available: http://chalearnlap.cvc.uab.es/
2. "2016 looking at people eccv challenge first impressions (first round)." [Online]. Available: http://chalearnlap.cvc.uab.es/challenge/14/track/14/description/
3. "2016 looking at people eccv challenge first impressions (second round)." [Online]. Available: http://chalearnlap.cvc.uab.es/challenge/15/track/15/description/
4. "2017 looking at people cvpr/ijcnn coopetition explainable impressions." [Online]. Available: http://chalearnlap.cvc.uab.es/challenge/23/track/22/description/
5. Xiu-Shen Wei,Chen-Lin Zhang,Hao Zhang, "Deep bimodal regression for apparent personality analysis," 2016.
6. Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, Anurag Mittal, "Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features," 2016.
7. Furkan Grpinar, Heysem Kaya, Albert Ali Salah, "Multimodal fusion of audio, scene, and face features for first impression estimation."
8. "Five video classification methods implemented in keras and tensorflow." [Online]. Available: https://blog.coast.ai/five-video-classification-methods-implemented-in-keras-and-tensorflow-99cad29cc0b5
9. "Opencv official webcite." [Online]. Available: https://opencv.org/
10. "Dlib official webcite." [Online]. Available: http://dlib.net/
11. Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, "Deep face recognition," 2015.
12. "Vgg-face pretrained model." [Online]. Available: robots.ox.ac.uk/ vgg/software/vgg_face/
13. Hochreiter, Sepp and Schmidhuber, Jurgen, "Long short-term memory," 1997.