# Automated Screening Of Job Candidate Based On Multimodal Video Processing

Jelena Gorbova
iCV Research Group
University of Tartu
Tartu 50411, Estonia
lena@icv.tuit.ut.ee

Iiris Lüsi
iCV Research Group
University of Tartu
Tartu 50411, Estonia
iiris@icv.tuit.ut.ee

Andre Litvin
iCV Research Group
University of Tartu
Tartu 50411, Estonia
andre@icv.tuit.ut.ee

Gholamreza Anbarjafari
iCV Research Group
University of Tartu
Tartu 50411, Estonia
shb@tuit.ut.ee

## Abstract

*The selection of adequate job candidates is very long and challenging process for each employer. The system presented in this paper is aiming to decrease the time for candidate selection on the pre-employment stage using automatic personality screening based on visual, audio and lexical cues from short video-clips. The system is build to predict candidate scores of 5 Big Personality Traits and to estimate a final decision, to which degree the person from video-clip has to be invited to the job interview. For each channel a set of relevant features is extracted, which are used to train separately from each other using Deep Learning. In the final stage all three results are fused together into final scores prediction. The experiment was conducted on first impression database and achieved significant performance.*

## 1. Introduction

It is well known-fact, that in choosing the right candidate, personality characteristics plays a relevant role on equal basis with candidate professional sills [1, 2]. It is still a challenge to automatically estimate professional skills of person, but with regard to personality analysis, there are successful studies designed to automatic personality analysis using different input data, e.g human speech, [3–9]. In [4] authors provide the model for automatic recognition of personality type from Mission Survival corpus video data set. In their work they use speech paralinguistic as well as social attention features. Social attention features, which are obtained by jointly processing the audio-video channels, were defined as the outcome of joint processing of head-pose and visual-gaze. The highest performance rate 0.59 they achieved with Naive Bayes classifier. The model based on linguistic text features is presented in [5]. Here authors provide a system for personality detection of online social networks users based on general perceptions. Unlike many other studies authors use here Three Factor Personality Model instead of 5 Big Personality Traits Model. In [3] authors apply a personality assessment paradigm to speech input to automatically estimate 5 Big Personality Trait scores. For speech analyze low-level speech features such as intensity, pitch,loudness, formants, MFCCs and Zero Crossing Rate are used. In this work results obtained with a model were compared with professional speaker estimation for the same parameters. The achieved result was about 60% in a ten class task consisting of isolated, acted productions of high and low targets for the 5 personality traits.

The most common approach in sociological field described the personality of human by basic traits, that are known as 5 Big Personality Traits [10]:

- **Extroversion** (sociability, assertiveness)
- **Agreeableness** to other people (friendliness)
- **Conscientiousness** (discipline)
- **Neuroticism** (emotional stability)
- **Openness** to experience (intellect)

In this paper, we present a system of automatic personality screening from short video presentations in order to make a decision whether a person has to be invited to a job interview. Considering the unstable situation with unemployment and steadying increasing competition in the world job market it's hard to imagine how many job candidates have to be reviewed by an employer in the first stage of candidates selection.

Drawing on the previous studies we take into account the assumption, that it's possible to implement the personality analysis without personal contact. Based on that we present a system aimed to estimate a persons scores in above mentioned personality characteristics as well as a estimate decision, whether the person has to be invited to the job interview, using short video clips. Our system benefits on multimodal processing methods [11]. We consider the visual factor (frames extracted from video), speech paralinguistic and lexical effect separately from each other. Our system provides more significant result within the human personality can't be based only on, for example, person's appearance. The importance of using multimodal processing was confirmed by the exceptional cases in the database (see Section 2).

In Section 2, we describe the first impression database, which was used for this work. In Section 3, we provide a full description of proposed system. In Section 4, the experimental results are presented.

## 2. Database description

The database used for this research was taken from challenge, which was organized within the Job Candidate Screening Competition Speed Interviews project. 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 [8, 12]. People in the videos belong to different age, gender, nationality and ethnic groups. Moreover the database represents some exceptional cases, e.g. on some of database videos people were speaking sign language, or there also cases, when person is sitting in front of camera without movement and saying a singe word. Each video is labeled with 6 values in range from 0 to 1. Five of them describe 5 big personality traits, namely extroversion, agreeableness, conscientiousness, neuroticism and openness. The Amazon Mechanical Turk (AMT) was used for measuring the traits values. With the $6^{th}$ (Interview) label AMT workers have estimated whether the person should be invited to a job interview or not. Since the research was done before the end of challenge annotations were announced only for training and validation set, only 8000 videos from first impression database were used in this work.

A pre-analysis was carried out on the provided data labels, had proved the assumption, that the final interview score can be estimated from 5 Big Personality Traits. For that we've explored the correlation between 'Interview' value and 5 Big Personality Traits estimations. For each of traits there's a high positive correlation with 'Interview' value (See Table 1).

Table 1. Correlation between 'Interview' value and 5 Big Personality Traits

| 5 BPT | Correlation |
|---|---|
| **Agreeableness** | 0.8228 |
| **Extroversion** | 0.8157 |
| **Neuroticism** | 0.8784 |
| **Openness** | 0.7694 |
| **Conscientiousness** | 0.8307 |

## 3. The proposed method

A block diagram of the proposed method is shown in Fig. 2. In our approach we consider each clip from database separately as speech signal, set of frames presented in video and set of words, which person uses when speaking. The impression, what a person leaves, plays the main role in personality analysis. As Mehrabian claims in his work [13] whether the listener feels liked or disliked depends only 7% on the spoken word, 38% on vocal utterances and 55% on facial expressions. The Fig. 1 illustrates very well the importance of visual effect in cases of job interview invitation as well as personality analysis. In Fig. 1 we show example frames from videos, which have extremal values in each database category.

In relation to automatic personality analysis through dominant paradigms it's impossible to take into consideration only the instance with the theoretically higher influence rate on like or dislike feeling. The person can looks friendly, smile a lot, visually project a nice picture, but on other hand the use of totally negative words in speech will adverse the decision, whether the person should be invited or not to a job interview. For each for these instances we extract features, select the subset of relevant ones (Section 3.1-3.3) and train them separately from each other(Section 3.4). After that we estimate the final decision by finding the optimal weights for predictions obtained using audio signal, video frames and content of speech (also known as text).

### 3.1. Speech paralinguistic features

Firstly the speech signal is extracted from video file and decoded into time series. Since in this work we are dealing with very short clips (around 15 sec), all of predicted values are mostly based on the first impression. In that case the personality traits and Interview values depend on that, what emotions a person express with his voice. Therefor, audio features, which are commonly used for automatic emotion recognition, could be used also for personality analysis and the Interview value estimation.

In emotion recognition field it has been shown, that the

Figure 1. Database examples with extreme values for each label

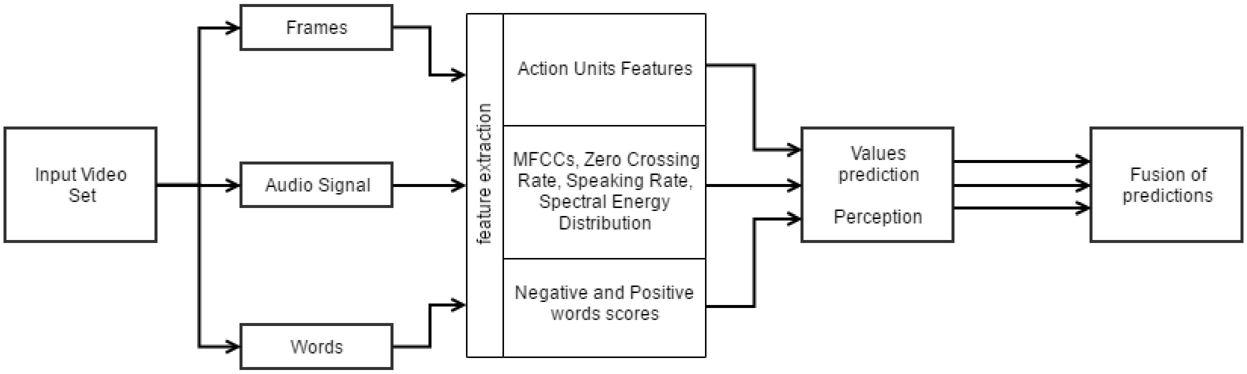

Figure 2. Block diagram of proposed method.

most relevant features for emotion recognition from human speech are pitch, intensity, duration, spectral energy distribution, Mel Frequency Cepstral Coefficients (MFCCs), formants, zero-crossing rate, and filter-bank energy parameters [14], [15]. In this work we use MFCCs, zero crossing rate, speaking rate and spectral energy distribution features (such as centroid, bandwidth and contrast). These are commonly applied not only in emotion recognition field, but also in some researches related to personality analysis [16], [3].

### 3.2. Video features

To extract and choose visual features we follow the same assumption as with audio features. From that, emotions, which person reflects, are the most relevant factor by speaker personality and 'Interview value' estimation relying the first impression .

For facial features extraction we use OpenFace[1], which provides a large number of facial features, namely 416 features for each of the frames. The set of features includes 2D facial landmarks location in pixels, 3D facial landmarks coordinates, head orientation in Euler angles, eye gaze vectors for left and right eye in world coordinates, the location of the head with respect to camera in millimeters, head rotation in radians, parameters of a point distribution model (PDM) that describe the rigid face shape and non-rigid face shape, Histogram of Oriented Gradients (HOG) features, intensity and presence of some of Facial Action Units (AUs), which OpenFace is able to detect.

In many instances people in provided videos are actively gesturing and changing their head pose, which decreases the precision of some features, e.g. facial landmarks, eye gaze direction. At the same time provided AUs features sufficiently characterize facial expressions, on which visual emotion recognition system is usually based. For that reason AUs features were chosen in this work as the most rel-

[1] free and open source face recognition with deep neural networks

Table 2. AUs features description

| AU | Description |
|---|---|
| **AU1** | Inner Brow Raiser |
| **AU2** | Outer Brow Raiser |
| **AU4** | Brow Lowerer |
| **AU5** | Upper Lid Raiser |
| **AU6** | Cheek Raiser |
| **AU7** | Lid Tightener |
| **AU9** | Nose Wrinkler |
| **AU10** | Upper Lip Raiser |
| **AU12** | Lip Corner Puller |
| **AU14** | Dimpler |
| **AU15** | Lip Corner Depressor |
| **AU17** | Chin Raiser |
| **AU20** | Lip stretcher |
| **AU23** | Lip Tightener |
| **AU28** | Lip Suck |

evant.

The list of detected in OpenFace AUs features is presented in the Table 2. For each of AUs in that list were detected intensity and presence, except for AU28, for which only the presence was detected. In total 29 features were extracted from each video frame.

On an average each video from provided database displays 28 frames per second. In order to reduce learning time and use only those frames, which would better represent a speaker, we choose 5 key frames from each video (see Fig. 3) using a clustering approach, which was successfully applied in several researches [17], [18].

Let us consider the cluster based key frames algorithm in the certain Video $V_i$. Let's say $X_i \in \mathbb{R}^{N^i \times 29}$ is a set of feature vectors, where $N^i$ is a number of frames in $V_i$. It is necessary to minimize the sum of squared errors, between the features vectors and their assigned centers:

$$J(X_i, C) = \sum_{j \in N_i} ||x(j) - C(a(j))||^2, \qquad (1)$$

where $C$ is set of centroids, $a$ is index of assigned center for each point [19]. To minimize (1) we use Lloyd's algorithm (see Algorithm 1).

In other words, this clustering-based strategy attempts to group frames with a similar posture. Each frame is assigned to a corresponding cluster, and those closest to the centroid of each cluster are selected as key-frames.

After key frames are chosen and extracted from video $V_i$, the new features vector is created by putting rows from key frames feature matrix $M_j \in \mathbb{R}^{5 \times 29}, j \in \{1, ..., 5\}$ in one vector one after the other.

**Algorithm 1** K-mean clustering

```
procedure LLOYD($(X_i, C)$)
    while not converged do
        for all $j \in N_i$ do Find the closest center to each $x(j)$
            $a(j) \leftarrow 1$
            for all $kinK$ do
                if $||x(j) - C(j)|| < ||x(j) - C(a(j))||$ then
                    $a(j) \leftarrow k$
        for all $kinK$ do Move the centers
            move $c(k)$ to the mean of $\{x(j)|a(j) = k\}$
```

### 3.3. Speech features

The first impression database provides a full text transcription for each video, which are obtained using a professional human transcription service text.

Let's say $S_i$ is vector of words speaker uses in speech. To get linguistic features from speech we use SentiWordNet[2], which provides negative and positive weights for more that 117000 synonyms sets. Each single word $s_i \in S_i$ was compared with SentiWordNet database. The corresponding positive and negative weights represent a new vector $W_i = \{[p_{s_i}, n_{s_i}] : i \in N_i\}$, where $p_{s_i}$ and $n_{s_i}$ are positive and negative weights of word $s_i$ from video $V_i$. A set of lexical features includes minimum, maximum, average and and sum of positive and negative weights from $W_i$. In total we get 8 features for each Video.

### 3.4. Regression model and combining the results

For value prediction we've applied multilayer Perceptron Neural Network (see Fig. 4) with Video, Audio Signal and speech content (text) features separately. On first stage perceptron computes net input value $z$ as the linear combination of feature variables $x$ and the model weights $w$

$$z = \sum x_i w_i. \qquad (2)$$

After that on $z$ value applied the threshold function

$$g(x) = \begin{cases} 1 & z \geq \Theta \\ -1 & otherwise \end{cases} \qquad (3)$$

where $\Theta$ is the threshold theta.

After we get prediction using separately audio, video and speech content (text) features, we have to combine them together. As far as visual, audio perception and the meaning of words the person uses, affect others perception in varying degrees, we calculate the weighted average, by using weights proposed in Mehrabian's work [13]. The final result is calculated by following formula:

$$P = 0.07P_t + 0.35P_a + 0.55P_v, \qquad (4)$$

[2] lexical resource for opinion mining

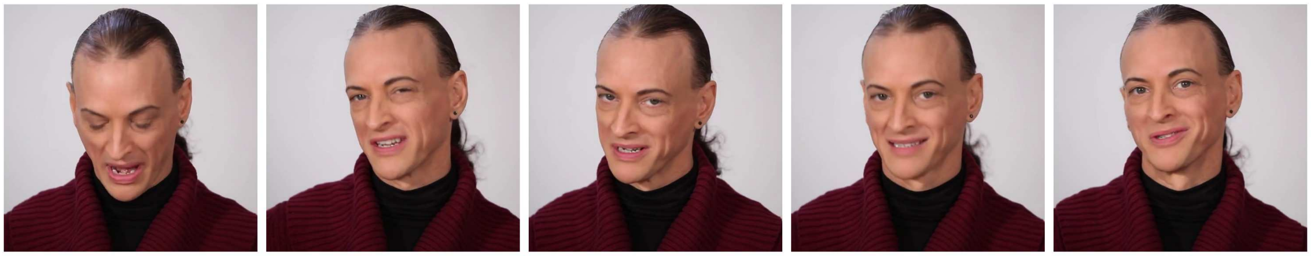

Figure 3. Extracted key frames using clustering approach based on Action Units features

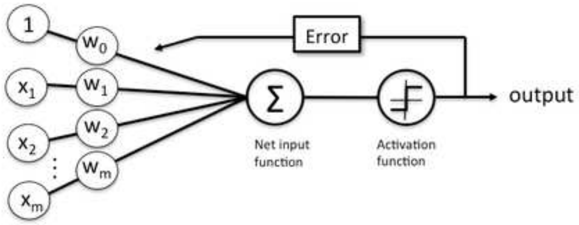

Figure 4. The perceptron algorithm

where $P_t, P_a, P_v$ are predictions obtained using speech content (text), audio and video features respectively.

## 4. The experimental results

In this section proposed method was applied on the first impression database. To obtain a broad picture of system efficiency the 8-fold validation was applied.

Firstly audio, video and speech content (text) features were extracted. For audio features extraction the Python library 'Librosa' was used, which is commonly used for audio and music analysis. For each audio file were extracted 40 MFCCs, Zero Crossing rate, speaking rate, mean of spectral centroid, bandwidth and contrast.

As it was described in proposed method section, for facial features extraction OpenFace was used. For clustering and key frames detection we've applied Matlab k-means function, with $k = 5$.

After feature extraction for all of three modes, the models were trained separately from each other. For training stage open source software Tensorflow was used, while it ensures relatively short learning time and provides a large set of functions used for Deep Learning. To predict 5 Big Personality Traits and Interview values we've applied multilayer perceptron with two hidden layers.

The performance rates in each category are presented in Table 3, which were calculated with following formula:

$$\frac{\sum_{N_t}(1 - |p_i - r_i|)}{N_f}, \qquad (5)$$

where $N_t$ is number of videos in validation set, $N_f$ total number of folds, $p_i$ and $r_i$ predicted and real values respectively.

As Table 3 shows the use of speech content (text) features provides the worst results out of all three channels used. In spite of that the difference between performance for each channel as well as for each value category is very small, the differences variate in range (0,02; 0.052) and the average for all audio, speech content (text) and video performance rates are 0.886, 0.885 and 0.886 respectively. The small differences allow us combine obtained results into the final decision.

For results fusion we calculate the weighted average, using the weights proposed by Mehrabian [13]. The final result is presented in the Table 3 in the $4^{th}$ column. As we see from Table 3 the average of performance rate for final predictions is higher than for audio, video and speech content (text) category. To justify, that the combination of all three channels for final prediction is better then the use only one of them, we analysed mean squared errors of obtained predictions for video, audio, speech content (text) and final category. The Table 4 present a standard deviation of mean square error (MSE) for each label. Only for 'Agreeableness' label the MSE was higher than the minimum in three other categories. At the same time the difference between them is only $0.02$. In other cases MSE is either lower or as low as the minimum of MSE for video, audio and speech content (text) predictions.

The same tendency occurs by analysis of MSE maximum. As we see in the Table 5 the maximum of MSE for final prediction is always the lowest in comparison to three other categories, except for 'Agreeableness' label.

On the basis of previous discussion the fusion of the predictions obtained by using video, audio and text features provides more significant and stable result with $89\%$ average performance for all of labels.

## 5. Conclusion

In this paper, a novel system for job candidate automatic screening using the short video-clips, which predicts 5 big personality scores of a person as well as estimates a decision, whether a person has to be invited to a job interview, was proposed. The final prediction in our system is

Table 3. Value predictions using video, audio, speech content (text) features and their final combination

| **Label** | **Audio** | **Text** | **Video** | **Final** |
|---|---|---|---|---|
| **agreeableness** | 0.904 | 0.901 | 0.896 | **0.902** |
| **conscientiousness** | 0.852 | 0.882 | 0.882 | **0.884** |
| **extraversion** | 0.89 | 0.874 | 0.886 | **0.892** |
| **openness** | 0.897 | 0.884 | 0.89 | **0.896** |
| **neuroticism** | 0.876 | 0.881 | 0.877 | **0.885** |
| **interview** | 0.895 | 0.888 | 0.887 | **0.894** |
| **avg** | 0.886 | 0.885 | 0.886 | **0.892** |

Table 4. Standard deviation of mean squared errors

| **Label** | **Audio** | **Text** | **Video** | **Final** |
|---|---|---|---|---|
| **agreeableness** | 0.021 | 0.024 | 0.025 | **0.023** |
| **conscientiousness** | 0.042 | 0.027 | 0.029 | **0.027** |
| **extraversion** | 0.022 | 0.022 | 0.027 | **0.022** |
| **openness** | 0.024 | 0.029 | 0.028 | **0.024** |
| **neuroticism** | 0.032 | 0.027 | 0.028 | **0.025** |
| **interview** | 0.021 | 0.026 | 0.026 | **0.021** |

Table 5. Maximum of mean squared errors

| **Label** | **Audio** | **Text** | **Video** | **Final** |
|---|---|---|---|---|
| **agreeableness** | 0.16 | 0.18 | 0.22 | **0.17** |
| **conscientiousness** | 0.32 | 0.21 | 0.21 | **0.18** |
| **extraversion** | 0.13 | 0.17 | 0.23 | **0.15** |
| **openness** | 0.21 | 0.18 | 0.2 | **0.18** |
| **neuroticism** | 0.22 | 0.23 | 0.19 | **0.18** |
| **interview** | 0.17 | 0.16 | 0.19 | **0.15** |

based on the combination of the results obtained by training video, audio and speech content (text) features. For values prediction the perceptron neural network was used in this work. The system has achieved the significant performance (in average for 6 labels 89%) by testing the system on first impression database.

## References

[1] C. Segalin, D. S. Cheng, and M. Cristani, "Social profiling through image understanding: Personality inference using convolutional neural networks," *Computer Vision and Image Understanding*, vol. 156, pp. 34–50, 2017.

[2] S.-J. Chen and L. Lin, "Modeling team member characteristics for the formation of a multifunctional team in concurrent engineering," *IEEE Transactions on Engineering Management*, vol. 51, no. 2, pp. 111–124, 2004.

[3] T. Polzehl, S. Moller, and F. Metze, "Automatically assessing personality from speech," in *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*. IEEE, 2010, pp. 134–140.

[4] J. Staiano, B. Lepri, R. Subramanian, N. Sebe, and F. Pianesi, "Automatic modeling of personality states in small group interactions," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 989–992.

[5] H. H. S. Sagadevan, N. Malim, "Sentiment valences for automatic personality detection of online social networks users using three factor model," *Procedia Computer Science*, pp. 201–208, 2015.

[6] B. Chen, S. Escalera, I. Guyon, V. Ponce-López, N. Shah, and M. O. Simón, "Overcoming calibration problems in pattern labeling with pairwise ratings: application to personality traits," in *Computer Vision–ECCV 2016 Workshops*. Springer, 2016, pp. 419–432.

[7] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language," *Journal of personality and social psychology*, vol. 108, no. 6, p. 934, 2015.

[8] J.-I. Biel and D. Gatica-Perez, "The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs," *IEEE Transactions on Multimedia*, vol. 15, no. 1, pp. 41–55, 2013.

[9] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 273–291, 2014.

[10] O. P. John and S. Srivastava, "The big five trait taxonomy: History, measurement, and theoretical perspectives," *Handbook of personality: Theory and research*, vol. 2, no. 1999, pp. 102–138, 1999.

[11] F. Noroozi, M. Marjanovic, A. Njegus, S. Escarela, and G. nbarjafari, "Fusion of classifier predictions for audio-visual emotion recognition," in *International Conference on Pattern Recognition (ICPR)*. Springer, 2016.

[12] J.-I. Biel, O. Aran, and D. Gatica-Perez, "You are known by how you vlog: Personality impressions and nonverbal behavior in youtube," in *ICWSM*, 2011.

[13] A. Mehrabian, "Communication without words," *Communication Theory,*, pp. 193–200, 2008.

[14] S. Haq and P. J. Jackson, "Multimodal emotion recognition," *Machine audition: principles, algorithms and systems*, pp. 398–423, 2010.

[15] D. Kamińska, T. Sapiński, and G. Anbarjafari, "Efficiency of chosen speech descriptors in relation to emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2017, no. 1, p. 3, 2017.

[16] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proceedings of the 2nd international workshop on Social signal processing*. ACM, 2010, pp. 17–20.

[17] Q. Zhang, S.-P. Yu, D.-S. Zhou, and X.-P. Wei, "An efficient method of key-frame extraction based on a cluster algorithm," *Journal of human kinetics*, vol. 39, no. 1, pp. 5–14, 2013.

[18] S. Hasebe, M. Nagumo, S. Muramatsu, and H. Kikuchi, "Video key frame selection by clustering wavelet coefficients," in *Signal Processing Conference, 2004 12th European*. IEEE, 2004, pp. 2303–2306.

[19] G. Hamerly and J. Drake, “Accelerating lloyd’s algorithm for k-means clustering,” in *Partitional clustering algorithms*. Springer, 2015, pp. 41–78.