Human-Explainable Features for Job Candidate Screening Prediction

Achmadnoer Sukma Wicaksana
Delft University of Technology
Delft, The Netherlands
a.s.wicaksana@student.tudelft.nl

Cynthia C. S. Liem
Delft University of Technology
Delft, The Netherlands
c.c.s.liem@tudelft.nl

Abstract

Video blogs (vlogs) are a popular media form for people to present themselves. In case a vlogger would be a job candidate, vlog content can be useful for automatically assessing the candidates traits, as well as potential interviewability. Using a dataset from the CVPR ChaLearn competition, we build a model predicting Big Five personality trait scores and interviewability of vloggers, explicitly targeting explainability of the system output to humans without technical background. We use human-explainable features as input, and a linear model for the systems building blocks. Four multimodal feature representations are constructed to capture facial expression, movement, and linguistic usage. For each, PCA is used for dimensionality reduction and simple linear regression for the predictive model. Our system's accuracy lies in the middle of the quantitative competition chart, while we can trace back the reasoning behind each score and generate a qualitative analysis report per video.

1. Introduction

Nowadays, content generated in the online world is largely based on multimedia. Each day, billions of hours of video are watched on YouTube and for each minute that passes by, three hundred hours of new videos are uploaded.[1] Included in this huge collection of content are video blogs (*vlogs*) that people use to present themselves and share anything to the world. The large amount of audiovisual vlog content has been shown to be useful for modeling and prediction of personality traits of the *vlogger* [4, 5].

In parallel, the popularity of video based content, combined with fast technology development, has also given rise to the video resume as a new type of job screening mechanism. For getting to know an applicant and understanding an applicant's hirability, the video resume offers advantages over a mere paper-based resume, which may suffer from judgmental bias [14]. Although not necessarily identical, vlog and video resumes have a similar form of one-way communication; the person speaks to the camera, and self-presentation will be an important motivation behind video production and sharing. The work at [18] tried to predict personalities and hirability from a video resumes dataset, and interest in this task also led to several ChaLearn 'Looking at People' benchmark challenges [19].

Unlike most of the machine learning problems that aim for only optimizing system accuracy, the problem of automatically assessing personality traits—and especially hirability of a potential job candidate—from audiovisual content needs to consider another aspect: explainability [21]. This type of work considers assessments of humans, and typical human decision-makers for this task do not have a technical computer science background. Therefore, it is critical to not only focus on numbers, but to also understand both measurements and the decisions made in a model. This means the chosen feature representations and model complexity have to be planned carefully.

In this paper, we present the work of using simple features that can be interpreted easily by a human to create a regression-based system for personality trait prediction and interviewability ('invite for interview') assessment. Using a simple linear regression model, the output of the system can be traced back to the original features, indicating how significant their influence is on the system. To support explainability of the automatic analysis, the system can generate text indicating a person's feature scores relative to others, and explaining these scores in relation to the final overall scoring of the model.

2. Human-Explainable Features

One important aspect that we have to keep in mind when predicting human personality scores is that the ground truth of assessment was done by a human, and that the final decisions on whether a person should be invited for a job interview will usually also be made by a human, who likely does not have a technical background. This means that the

[1] https://www.youtube.com/yt/press/statistics.html

model has to be as transparent and explainable as possible.

Considering this, our proposed model employs features can easily be described in natural language, employs a linear (PCA) transformation to reduce dimensionality, and uses simple linear regression models for predicting scores, such that scores can be traced back to and justified with the underlying features. While using hand-crafted features and models of this simplicity are not what typically is seen in state-of-the-art automatic solutions, we consider this explainability a clear strength. As we will demonstrate, within the ChaLearn benchmarking campaign, our model indeed is not the strongest in the quantitative sense, but not the weakest either, and human-readable descriptions generated from it are well appreciated by human judges.

The model considers two modalities, visual and textual, for extracting features. In the visual modality, we consider features capturing facial movement and expression, as they are one of the best indicators for personality [17, 7].

However, considering findings in organizational psychology, personality traits are not the only (and neither the strongest) predictors for job suitability. In fact, GMA (General Mental Ability) tests, such as intelligence tests, have the highest validity at the lowest application cost [20, 9]. While we do not have formal GMA assessments for subjects in our dataset, we consider that language use of the vlogger may indirectly reveal GMA characteristics, such as use of difficult words. This is why we also consider textual features, both considering speaking density, as well as linguistic sophistication.

2.1. Visual Features

For the visual representation, the system was not built to focus on the video in general, but particularly on facial expression and movement. In order to do this, we used OpenFace tools to segment only the face from each video, standardizing the segmented facial video to 112x112 pixels. OpenFace is an open source toolkit which does not only segment faces, but offers a feature extraction library that can extract and characterize facial movements and gaze [3]. OpenFace is able to recognize a subset of individual Action Units (AU) that construct facial expressions encoded in Facial Action Code System (FACS) as shown in Table 1 [10, 11]. These AUs then can be described in two ways: in terms of presence (indicating whether a certain AU is detected in a given time frame) and intensity (indicating how intense an AU is at a given time frame).

For each of these AUs, we construct three features for our input to the system. First, we consider the percentage of time frames during which the AU was visible in a video. Second, we store the maximum intensity of the AU in the video. Lastly, we also store the mean intensity of the AU over the video. These three features per AU add up to 52 features in total for the OpenFace representation.

Action Unit	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
AU25	Lips part
AU26	Jaw Drop
AU28	Lip Suck
AU45	Blink

Table 1. Action Units that are recognized by OpenFace and its description

The resulting face segmented video also is used for another video representation. In order to capture overall movement of the vlogger's face, a Weighted Motion Energy Image (wMEI) is constructed from the resulting face segmented video. MEI is a grayscale image that shows how much movement happens on each pixel throughout video, with white indicating a lot of movement and black indicating less movement [6]. wMEI was proposed in the work by Biel *et al*. [4] as a normalized version of MEI, by dividing each pixel values with the maximum pixel value. Our method is inspired by the aforementioned work with improvement on background noise reduction. In the said work, the whole video frame is used as an input to compute wMEI, which makes background movement contribute to the overall wMEI measurements. Thus, there are cases in which the resulting wMEI is all white due to background movement, rather than movement of a human subject. For example, this happens when the vlogger recorded the video in a public space or while on the road. By departing from our face segmented video instead of a whole video frame, we minimize the involvement of background in our calculation and thus get a better representation of the subject's true movement, as we can see in Figure 1. In order to create wMEI, we obtain the base face image of each video and iterate over the video time frames to compute the overall movement for each pixels. For each wMEI, three statistical features (mean, median, and entropy) are extracted to constitute a MEI representation.

The current dataset [19] that we are working on for this problem has been carefully selected so that only one unique

Figure 1. wMEI for face segmented video

foreground person faces the camera in the video. However, the current OpenFace implementation has limitations when the video still contains other visual sources with faces, such as posters or music covers in the background. While the situation is rare, we occasionally noticed that a poster was detected and segmented as 'main face' rather than the subject's actual face. For such misdetections, no movement will be detected at all, so this corner case is easily captured by our system and reported on in our feature description.

2.2. Textual Features

Textual features are generated by using transcripts that were provided as the extension of the [19] dataset. For a handful of videos, transcript data was missing; we manually annotated those videos, such that all videos have transcript data for our purposes, with exception of one video that has no transcript because the person speaks in sign language in the video.

As reported in literature [20, 9] and confirmed in private discussions we had with organizational psychologists, assessment of GMA (intelligence, cognitive ability) is important for many hiring decisions. While this information is not reflected in personality traits, we felt that the linguistic usage of the subjects may possibly reveal some related information.

To assess the linguistic usage of the vlogger, we employed several Readability indexes on the transcripts. This was done by using open source implementations of various readability measures in the NLTK-contrib package of the Natural Language Toolkit (NLTK). More specifically, we used 8 measures as features for the Readability representation: ARI [22], Flesch Reading Ease [12], Flesch-Kincaid Grade Level [15], Gunning Fog Index [13], SMOG Index [16], Coleman Liau Index [8], LIX, and RIX [2]. While these measures are originally developed for written text (and officially may need longer textual input than a few sentences in a transcript), our expectation still would be that they would consistently reflect complexity in linguistic usage. In addition, we also used two simple statistical features for an overall Text representation: total word count in the transcript, and the amount of unique words within the transcript.

3. Traits Prediction

The building blocks of our predictive model encompass four feature representations; OpenFace, MEI, Readability, and Text. Employing the 6000 training set videos, for each representation, we train a separate model to predict personality traits and interview scores. For a final prediction score, we apply late fusion and average the predictions made by the four different models. A diagram of the proposed system can be seen at Figure 2.

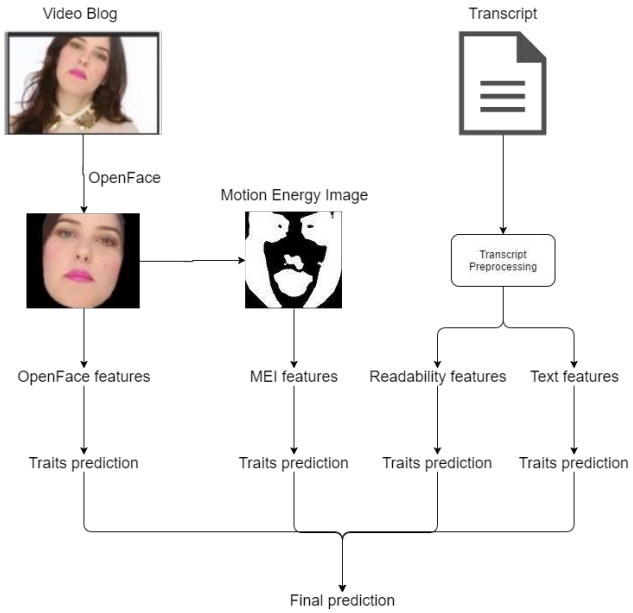

Figure 2. Overall system diagram

As the goal of our system is to trace back the prediction scores to each underlying features, linear models are best suited for our purposes. It also should be noted that linear regression is a commonly seen model in social sciences literature. We apply Principal Component Analysis (PCA), which performs an orthogonal linear transformation of features, as dimensionality reduction technique for each representation, retaining 90% variance. The resulting transformed features then act as input for a simple linear regression model to predict the scores.

As we only use linear models, we can trace the importance of each original feature for our final prediction. By considering the linear regression coefficients, we know for each PCA dimension whether it contributes positively or negatively to the prediction. Furthermore, considering the PCA transformation matrix, we can trace back how strongly each original feature contributed to each PCA dimension.

While we did not formally participate in the quantitative phase of the ChaLearn CVPR2017 competition[2], Table 2

[2] http://chalearnlap.cvc.uab.es/challenge/23/description/

Categories	Our System	Lowest	Highest
Interview	0.887744	0.872129	0.920916
Agreeableness	0.896825	0.891004	0.913731
Conscientiousness	0.880077	0.865975	0.919769
Extraversion	0.887040	0.878842	0.921289
Neuroticism	0.884847	0.863237	0.914613
Openness	0.890314	0.874761	0.917014

Table 2. Accuracy (1 - Mean Absolute Error) comparison between our proposed system and the lowest and highest accuracy for each prediction category in the ChaLearn CVPR 2017 Quantitative challenge.

shows the overall quantitative accuracy of our system on the 2000 videos in the benchmark training set, for each of the Big Five personality traits and the interview invitation assessment. For each predicted class, we compare our scores to the lowest and highest scores (from all of the participants) in the ChaLearn CVPR 2017 Quantitative challenge.

As expected, our system does not yield optimal accuracy, but comparing our scores to the officially reported scores in the Quantitative challenge, our proposed system would consistently outperform the lowest-scoring system for each category. This comes with the benefits of low computational power demands for model fitting, and the earlier discussed advantages for explainability due to our linear models.

4. Explainability

In the Qualitative phase of the ChaLearn CVPR 2017 challenge, the goal was to explain predictions by a human-understandable text. We implemented a simple text description generator, departing from the following thoughts:

- As explained before, each of our visual and textual features were picked to be explainable in natural language to a non-technical human;
- However, we do not have any formal proof which of our features are fully valid predictors of personality traits or interviewability. While our model gives indicators on the strongest linear coefficients, the assessments it was trained on are made by external observers (crowdsourcing workers), which poses a very different situation from the assessment settings in the formal psychology studies as reported in [20]. Therefore, we will not make a hard choice for 'good' features yet, but rather provide a comprehensive report on each observed feature, also indicating acknowledgement of potential feature weaknesses (e.g. indicating that several readability scores were developed for larger texts);
- It may be possible to aggregate feature observations to higher-level descriptions (in particular, regarding AU detections, as combinations of AUs may indicate higher-level emotional expressions), but as this would increase the complexity of our model, we will for now maintain a basic description describing individual low-level features;
- As our feature measurements did not formally get tested yet in terms of psychometric validity, it is debatable to consider feature measurements and predicted scores as absolute indicators of interviewability. However, for each person, we can indicate whether the person scores 'unusually' with respect to a larger population of 'representative subjects' (formed by the vloggers represented in the 6000-video training set). Therefore, for each feature measurement, we report what the typical range of the feature is, and at what percentile the feature score of the subject is, compared to scores of the subjects in the training set.
- Finally, to still reflect major indicators from our linear model in our description, for each representation (OpenFace, MEI, Readability, Text) we pick the two linear regression coefficients that are largest in the absolute sense. For the PCA dimensions corresponding to these coefficients, we trace back which two features contributed most strongly to this PCA dimension, and whether the features contribute positively or negatively to the linear model. For these features, a short notice is added to the description, expressing how the feature commonly affects final scoring (e.g. 'In our model, a higher score on this feature typically leads to a higher overall assessment score' for a positive linear contribution.

As a result, for each video in the validation and test set, a fairly long but consistent textual description was generated. An example fragment of the description are given in Figure 3.

As part of the ChaLearn competition challenge, our descriptions were evaluated by a human jury. The corresponding scores are reported in Table 3. While our average scores were slightly lower than that of the other submitted system in the challenges, our system led to higher standard deviations (possibly indicating stronger jury responses), and ultimately the differences between the systems were not deemed statistically significant.

Evaluation	Scores
Clarity	3.33±1.43
Explainability	3.23±0.87
Soundness	3.43±0.92
Interpretability	2.4±1.02
Creativity	3.4±0.8

Table 3. Explainability scores

```
******************
* USE OF LANGUAGE *
******************

Here is the report on the person's language use:


** FEATURES OBTAINED FROM SIMPLE TEXT ANALYSIS **
Cognitive capability may be important for the job.  I looked at a few very simple
text statistics first.

*** Amount of spoken words ***
This feature typically ranges between 0.000000 and 90.000000.  The score for this
video is 47.000000 (percentile:  62).
In our model, a higher score on this feature typically leads to a higher overall
assessment score.
```

Figure 3. Example description fragment.

5. Conclusions and Future Work

We presented a system for personality trait and interviewability prediction, which was designed such that the system's underlying features and decision-making processes were as transparent as possible. Despite the simplicity of our features and models, reasonable quantitative system accuracy scores were obtained. Qualitative natural language descriptions generated from our model also were judged positively by the jury members in the challenge.

To maintain explainability, for future improvements to our system, our preference would be to keep using linear prediction models. At the same time, several improvements can be performed, as explained below.

For the feature generation phase, other human-explainable feature representations can be added to the system to generate improved accuracy. For example, audio features such as proposed in [4] can be integrated, which can help the system to further characterize the vlogger, and especially speaking characteristics. In addition to the current OpenFace representation 'as is', we could combine different AUs to assess higher-level emotions, such as Joy, which is typically characterized as a combination of AU12, AU6, and AU7 [23, 11]. While this will increase the complexity of the system, it is likely that emotions are better-interpretable to humans than individual AUs.

The current system currently applies late fusion, and presents the final prediction as the average between four types of feature representations. Instead of giving equal weight to each representation, we may be able to obtain better results when weighting each representation differently. Such further tuning can be performed by more extensive cross-validation on the training and validation datasets. Since our aggregation is linear, our system can incorporate such weight adjustments easily.

As for our textual descriptions, we currently did not select any strong features, but provided a full report on every single feature. This may have made our current report somewhat long and overwhelming to a human reader. We expect that our explanations will allow for a better user experience when presented in a less textual way, e.g. in the form of graphs. Next to this, to avoid information overload, smarter information selection can be performed. However, in order to do this, it is important to validate our features more strongly in the psychometric sense, and it will be useful to obtain further qualitative input from human judges on what parts of our explanation were understood and appreciated, and what parts were deemed less interpretable. It will particularly be important to receive such feedback from organizational psychologists and HR specialists, as those will be the most likely users and final decision-makers for a system like ours.

Lastly, it should be noted that the current dataset considers vlogs, but not official video resumes. Although there are similarities between vlogs and video resumes, video resumes might have distinct differences in term of content delivery and preparation [18]. For example, it is safe to assume that when people want to apply for a job, they will want to maximally impress a potential employer, rather than presenting themselves casually and more spontaneously, which may be the case in vlogs. Furthermore, the vlogs also were not targeted at a specific job vacancy, while job-specific demands may in reality be important for candidate assessment. It would therefore be useful to also acquire data on more realistic video resumes, even while this may negatively affect data scale. Finally, it will be useful to augment the current dataset with annotations beyond personality traits and interviewability, e.g. also testing for indicators of general mental ability.

For reproducibility, the code used in our submission to the ChaLearn challenge is made available on GitHub[3].

[3] https://github.com/sukmawicaksana/CVPR2017

References

[1] H. Abdi and L. J. Williams. Principal component analysis, 2010.

[2] J. Anderson. Lix and Rix: Variations on a Little-known Readability Index. *Journal of Reading*, 26(6):490–496, 1983.

[3] T. Baltrušaitis, M. Mahmoud, and P. Robinson. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *FG*, volume 06, pages 1–6, 2015.

[4] J.-I. Biel, O. Aran, and D. Gatica-Perez. You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube. *Artificial Intelligence*, pages 446–449, 2011.

[5] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez. Hi YouTube! Personality impressions and verbal content in social video. *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*, pages 119–126, 2013.

[6] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3):257–267, 2001.

[7] P. Borkenau, S. Brecke, C. Möttig, and M. Paelecke. Extraversion is accurately perceived after a 50-ms exposure to a face. *Journal of Research in Personality*, 43(4):703–706, 2009.

[8] M. Coleman and T. L. Liau. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975.

[9] M. Cook. *Personnel Selection: Adding Value Through People*. Wiley-Blackwell, fifth edition, 2009.

[10] P. Ekman and W. V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. 1978.

[11] P. Ekman and E. Rosenberg. *What the face reveals*. 2005.

[12] R. Flesch. A New Readability Yardstick. *The Journal of Applied Psychology*, 32(3):221–233, 1948.

[13] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952.

[14] A. M. F. Hiemstra. *Fairness in Paper and Video Resume Screening*. PhD thesis, Erasmus University Rotterdam, the Netherlands, 2013.

[15] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Technical Training*, Research B(February):49, 1975.

[16] G. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639646, 1969.

[17] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling. Personality judgments based on physical appearance. *Personality and social psychology bulletin*, 35(12):1661–1671, 2009.

[18] L. S. Nguyen and D. Gatica-Perez. Hirability in the Wild: Analysis of Online Conversational Video Resumes. *IEEE Transactions on Multimedia*, 18(7):1422–1437, 2016.

[19] V. Ponce-López, B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. Escalante, and S. Escalera. Chalearn LAP 2016: First round challenge on first impressions - Dataset and results. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9915 LNCS(November), 2016.

[20] F. L. Schmidt and J. E. Hunter. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological bulletin*, 124(2):262–274, 1998.

[21] G. Shmueli. To explain or to predict? *Statistical Science*, 25:289–310, 2010.

[22] E. A. Smith and R. J. Senter. Automated readability index. *AMRL-TR. Aerospace Medical Research Laboratories (6570th)*, pages 1–14, 1967.

[23] Y. L. Tian, T. Kanade, and J. F. Conn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001.