FaceTube: Predicting Personality from Facial Expressions of Emotion in Online Conversational Video

Joan-Isaac Biel
Idiap Research Institute
Ecole Polytechnique Fédérale
de Lausanne (EPFL)
jibiel@idiap.ch

Lucía Teijeiro-Mosquera
University of Vigo
Idiap Research Institute
luciatm@gts.uvigo.es

Daniel Gatica-Perez
Idiap Research Institute
Ecole Polytechnique Fédérale
de Lausanne (EPFL)
gatica@idiap.ch

ABSTRACT

The advances in automatic facial expression recognition make possible to mine and characterize large amounts of data, opening a wide research domain on behavioral understanding. In this paper, we leverage the use of a state-of-the-art facial expression recognition technology to characterize users of a popular type of online social video, conversational vlogs. First, we propose the use of several activity cues to characterize vloggers based on frame-by-frame estimates of facial expressions of emotion. Then, we present results for the task of automatically predicting vloggers' personality impressions using facial expressions and the Big-Five traits. Our results are promising, specially for the case of the Extraversion impression, and in addition our work poses interesting questions regarding the representation of multiple natural facial expressions occurring in conversational video.

Keywords

YouTube, vlogs, facial expressions, personality prediction

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]

1. INTRODUCTION

Conversational vlogging is a successful genre of online social video that generates an enormous amount of audiovisual behavioral content. One differential feature of vlogging is that users can combine their verbal expression of opinions, desires, and personal narratives, together with a myriad of spontaneous nonverbal cues (through face, body, and voice) captured by the audio and video channels, and all this is watched by their online audience. This people-watching-people phenomenon involves complex interpersonal perception processes, in which people watching build impressions from personal information inferred from vlogs, and based on them, may react by posting comments or sharing the videos with others. This poses questions regarding the types of impressions that are built, their reliability, and the sources of information that people use, which are important issues not only for the understanding of the production and consumption mechanisms of this social media, but also for the automatic characterization of vloggers and the prediction of interpersonal impressions based on the automatic analysis of vlog content. Due to their abundance, the variety of topics and the naturalistic conditions of vlogs we posit that facial expressions of emotion occur frequently enough (see Fig. 1) to influence personality judgements [10].

Recent research has addressed the study of interpersonal perception in vlogging, from the perspective of Big-Five personality impressions and nonverbal behavioral analysis [3, 4]. Regarding judgements of personality, Biel et al. found that Extraversion and Agreeableness are the traits judged with higher reliability in vlogging [3]. Moreover, in an attempt to address the personality prediction task from automatically extracted nonverbal cues [4], the same authors found that nonverbal cues from audio and visual activity patterns seemed useful mainly to predict Extraversion [4]. Compared to past attempts to predict personality meetings [11] and monologue video presentations [2], the results in [3, 4] emphasized that the cues conveying personality information and the specific impression that can be reliably estimated from automatic analysis are particular of each communication scenario. In addition, those authors did not investigate other possible sources of useful personal information. In other related work, the analysis of facial expressions on online video has been previously researched in [13] from the perspective of passive, mainly silent, viewing of advertising content, but to our knowledge our paper is the first attempt to deal with facial expressions in video vlogs where people are mainly talking. We address the problem of predicting vloggers personality impressions from automatically extracted facial expressions of emotion. The human face has been widely documented in the social psychology literature as an important source of information in interpersonal impressions [9, 10, 8]. People rely heavily in facial cues to make interpersonal judgements because there is a general belief that faces provide valuable information about a person's character or personality [9]. We argue that this may be specially true in the vlogging scenario, in which vloggers typically display head and shoulders in camera and faces occupy a large portion of the screen [4]. Among facial features, there is evidence that facial expressions of emotion provide information other that emotional states, influencing interpersonal impressions such as personality judgements, and that specific affective cues are in fact correlated with the possession of various personality traits [10, 8].

This paper has four contributions. First, we explore to what extent spontaneous facial expressions are found in vlogging based on a dataset of vlogs collected from YouTube

ICMI'12, October 22–26, 2012, Santa Monica, California, USA.

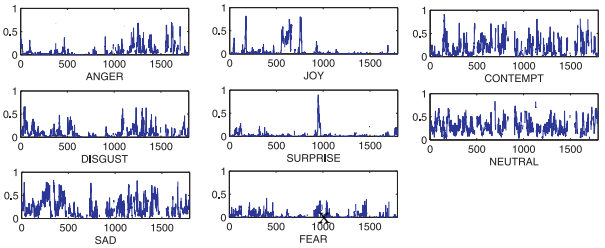

Figure 1: Example of CERT outputs for the eight universal facial expression of emotion.

and the analysis of the output obtained from processing vlogs with the Computer Expression Recognition Toolbox (CERT) [12]. Second, we study the associations between automatically extracted facial expressions of emotion and Big-Five personality judgements from vloggers to understand what specific facial expressions are most prominent for modeling each of the different impressions. Third, we do some experiments on predicting personality impressions usign different sets of facial expressions cues. Finally, we compare our results with previous results on personality prediction using nonverbal cues.

2. DATASET

Our dataset of vlogs was obtained from YouTube with a keyword-based search for "vlogs" and "vlogging", similarly to what was proposed in [4]. From this search we manually selected videos that featured a monologue scenario in which users talk in front of the camera. Then, we used a Viola-Jones face and facial feature detector to detect face, eyes, nose, and mouth [1], and selected videos were most of the frames contained all facial features detected. This preprocessing step for selecting videos aimed to minimize possible problems in the face registration step previous to the facial expression extraction (see Sec. 3.1). The final dataset contained 281 vloggers, mostly balanced in gender (53% female and 47% males). All the videos were cut to be one minute long.

We completed our dataset with annotations of personality impressions using Mechanical Turk. The task of crowdsourcing personality impressions from video watching had been attempted in previous work with comparable reliabilities to those obtained in other settings [4]. Using a short personality questionnaire [7], annotators were asked to judge the extent to which each vlogger could be described with each of the Big-Five traits: Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotion Stability (ES), and Openness to Experience (O). To have more reliable estimates for personality we collected five different annotations for each vlog and aggregated the personality impressions using the average value. The ICC(1,k) reliabilities for each one of the personality impressions were .76 (E), .63 (A) .42 (C), .40 (ES), and .49 (O).

3. AUTOMATIC EXTRACTION OF FACIAL EXPRESSION CUES

In this section, we explain our approach to characterize vloggers' facial activity using the frame-by-frame estimates of a state-of-the art facial expression recognition system.

3.1 Automatic Facial Expression Analysis

Facial expression analysis has been thoroughly researched during the last two decades [6, 5] in the computer vision field. The Facial Action Coding System (FACS) developed by Ekman et al. [5] has become a standard framework for detecting facial actions and for classifying facial expressions of emotion. FACS defines the action units (AUs) that code the movement of facial muscles, and are considered the fundamental units of facial expressions. Using FACS, seven facial expressions of emotion considered as universal - Anger, Contempt, Surprise, Fear, Joy, Sad and Disgust- are uniquely defined in terms of AUs.

We processed vlogs using the Computer Expression Recognition Toolbox (CERT) [12], which is a real-time face processing software developed for facial expression understanding and constitutes a state-of-the-art in the field. CERT combines three face processing stages: face detection, feature detection, and face registration, to obtain a cropped face patch used for expression analysis. Our video selection step based on facial feature detection was targeted to prevent CERT from mis-registrations. Based on the AUs estimated using Gabor-based filters and SVM classifiers, CERT uses a multivariate logistic regressor to predict the seven expressions of emotion plus neutral expression. We use these 8 categories for the purpose of characterizing the facial expressions. In addition, we also use the smile detector output present in CERT [12].

3.2 Facial expressions activity cues

We propose a systematic method to model the frame-by-frame facial expression stream from CERT (see Fig. 1) into aggregate cues that capture different facial expression activity patterns. First, we converted the CERT output to a binary segmentation that divides the expression signal in active/inactive regions using two different approaches:

Thresholding (THR): we consider the CERT output to be activated when frame values are larger than a threshold λ. For facial emotions we choose $\lambda = .005$, which represents a rather conservative choice, whereas for smile the threshold was set up to $\lambda = 0$ by definition [12].

HMM: we used a two-state (active/inactive) Hidden Markov Model to detect the active state of the CERT output. Each output is modeled with one single Gaussian initialized with the threshold-based segmentation, while the transition probabilities are set to $\rho_{00} = \rho_{11} = 0.95$ and $\rho_{01} = \rho_{10}$.

In practice, the THR approach copes with high frequency changes, tends to give shorter and more frequent active states, and is also more sensitive to outliers. The HMM provides a smooth output, that tends to detect peaks in the CERT generated signals.

Based on these segmentations we extracted four different facial cues to measure the presence of expressions, their duration, their frequency, and the time of short segments.

Proportion of active time (PT): computed as $PT = \frac{1}{N}\sum_{i=0}^{N_r}\tau(r_i = 1)$, where $\tau(r)$ is the duration of segment r in frames, N_r is the total number of segments, and N is the total number of frames.

Number of active segments (NS): computed as $NS = \frac{1}{Nf}\sum_{i=0}^{N_r}(r_i = 1)$, where f is the frame rate.

Average duration of the expressions (AD): computed as $AD = \frac{1}{Nf}\sum_{i=0}^{N_r}\tau(r_i = 1)$ where $\tau(r_i)$ is the duration of region r in frames.

Proportion of short segments time (PTS): computed as $PTS = \frac{1}{N_r}\sum_{i=0}^{N}\tau(r_i = 1|\tau(r_i) \leq 0.001f)$, i.e., the proportion of time in segments shorter than 100ms.

	THR	HMM
E	**Anger:** PT (−.16), AD (.24), NS (−.23), PTS (.27). **Contempt:** NS (−.15). **Disgust:** PT (−.11), AD (.22), PTS (.29). **Fear:** PT (.22). **Joy:** PT (.20), AD (.21), PTS (.17). **Neutral:** NS (−.18). **Surprise:** PT (.15), AD (.11). **Smile:** PT (.25), AD (.21), PTS (.18) # Cue utilization = 18	**Anger:** PT (−.12), AD (−.13). **Contempt:** PT (−.12), AD (−.21), NS (.21), PTS (.18). **Disgust:** PT (−.07). **Fear**: NS (.30), PTS (.26), **Joy:** NS (.37). **Neutral:** (AD (−.16), NS (.17). **Sad:** NS (.23), PTS (.13). **Surprise:** NS (.26), PTS (.24). **Smile:** PT (.19), NS (.22). # Cue utilization = 18
C	**Contempt:** NS (.10), **Joy:** PT (.10), NS (.09), PTS (−.10). **Neutral:** AD (−.19), NS (.20). **Surprise:** AD (−.11), PTS (−.13). # Cue utilization = 8	- # Cue utilization = 0
O	**Anger:** PT (−.22), AD (.11), NS (−.21), PTS (.16). **Disgust:** PT (−.14), AD (.10), PTS (.14). **Fear:** PT (.22), **Joy:** PT (.12). **Surprise:** PT (.20), NS (.20) # Cue utilization = 11	**Anger:** PT (−.14). **Disgust:** NS (−.10). **Fear:** NS (.23), PTS (.19). **Smile:** PT (.10), NS (.11). **Surprise:** NS (.21), PTS (.15). # Cue utilization = 8
A	**Anger:** PT (−.17). **Joy:** PT (.17), NS (.19), PTS (−.12). **Smile:** PT (.18), AD (.11). # Cue utilization = 7	**Anger:** NS (−.19). **Disgust:** NS (−.08), PTS (−.08). **Joy:** PT (.22), **Sad:** PTS (−.09). **Smile:** AD (.14). # Cue utilization = 6
ES	**Joy:** AD (−.16), NS (.09), PTS (−14). **Smile:** PT (.07). # Cue utilization = 4	**Anger:** NS (−.13). **Disgust:** NS (−.05), PTS (−.07). **Sad:** PTS (−.11). **Smile:** AD (.08). # Cue utilization = 5

Table 1: Significant Pearson correlation effects ($p < .05$.) between Big Five personality impressions (E, C, O, A, ES) and facial expression activity cues measured from THR and HMM segmentations.

	Thr			*HMM*		
Expression	Med	SD	Q_3	Med	SD	Q_3
Anger	.47	.25	.67	.02	.21	.06
Contempt	.88	.17	.95	.42	.41	1.00
Disgust	.33	.27	.62	.01	.14	.04
Fear	.50	.26	.68	.03	.17	.09
Joy	.54	.28	.75	.03	.24	.10
Neutral	.93	.14	.98	1.00	.23	1.00
Sad	.86	.16	.94	.20	.41	1.00
Surprise	.86	.16	.94	.20	.41	1.00
Smile	.12	.15	.26	.32	.15	.41

Table 2: Median and SD values for the PT cues obtained from Thr and HMM segmentations

This activity cue was introduced to explore the characterization of facial expressions as signals of short duration.

4. RESULTS

This section is divided in four parts. First, we provide descriptive statistics of vloggers' facial expression activity. Second, we analyze the correlation between facial expressions and personality impressions. Third, we address the task of personality prediction. Finally, we compare our results to other nonverbal cues reported in previous work.

4.1 Descriptive analysis

We computed basic statistics of facial activity cues across vlogs, but for space reasons in Table 2 we only report values of the PT cues obtained from both THR and HMM segmentations (see Table 2). These measures are useful to interpret the functioning of the two approaches proposed, and have also potential for understanding the type of facial expressions that can be typically found in the vlogging scenario.

Though THR and HMM segmentations provide substantially different values for most expressions, they seem to agree on the high presence of the Neutral expression: in Table 2, around half of the vloggers seem to have neutral expressions between 93 and 100% of the time. For the THR segmentation, this result challenges the fact that, as seen from the median values, a large number of the vlogs also show large presence of other facial expressions, which indicates that the intervals in which facial expressions are active overlap. Though it may seem the case by looking at the CERT output (see Figure 1), it is unlikely that some facial expressions co-occur in time. In contrast, the HMM segmentation suggests a more realistic scenario in which the activation of facial expressions signals concurs with the activation of the neutral signal. Though median values for HMM may seem low compared to THR segmentations, the third quartile (Q_3) indicates that 25% of the vlogs still show large presence of facial expressions of emotion such as Contempt, Joy, Sadness, Surprise, and Smiles.

4.2 Correlation analysis

We now investigate the individual correlations between facial expression activity cues and personality impressions, to understand what facial expressions may be useful to infer personality judgements in vlogging (a.k.a. cue utilization). Table 1 shows the significant effects ($p<0.05$) found for cues from both THR and HMM segmentations.

As a first result, we found that Extraversion is the trait showing the largest cue utilization (18), which indicates that facial expressions of emotion are useful to build personality impressions for this trait, which is related to the evidence showed in related literature that Extraversion is typically easier to judge [3, 4]. Though Agreeableness has the second largest ICC (see Sec. 2), it does not compare to Extraversion in terms of cue utilization (6-7), and it is even lower than other traits such as Openness to Experience.

We found different interesting effects between facial expressions of emotions and personality impressions. Facial expressions of emotions with positive valence such as Joy and Smile showed almost exclusively positive effects with personality impressions. For the case of HMM, for example, smile is correlated with Extraversion (PT, $r = .19.$), and Agreeableness (AD, $r = .14$), among others. Facial expressions of negative valence, such as Anger, showed negative correlations with judgements of Extraversion (PT, $r = -.16.$), Openness to Experience (PT, $r = -.22.$), and Agreeableness (PT, $r = -.17.$). Whereas Contempt or Surprise have both negative and positive effects, see for example Contempt and Extraversion impressions (NS, $r = -.15.$) compared to Conscientiousness (NS, $r = .10$).

We also saw differences between effects in cues computed from THR and HMM segmentations. Whereas most values of PT show same sign effects, some values for AD, NS, and PTS indicate different sign effects. Though this is clearly

Feature Set	E	C	OE	A	ES
THR	.17	.07	.12	.06	.05
HMM	.16	.07	.12	.06	.05
THR+HMM	.20	.05	.09	.07	.03
THR_Sel	.20	.07	.10	.06	.07
HMM_Sel	.16	-	.07	.05	.05
THR_Sel+HMM_Sel	.22	.06	.08	.07	.03

Table 3: Big-Five personality traits prediction results (in mean R-squared) for feature sets obtained from segmentations using THR and HMMs.

Feature Set	E	C	OE	A	ES
Audio	.37	.05	.06	.05	.02
Visual	.13	.04	.05	.02	.03
THR_Sel+HMM_Sel	.22	.06	.08	.07	.03

Table 4: Big-Five personality traits prediction results (in mean R-squared) for audio, visual, multimodal, and facial expression features.

consecuence of the two different segmentation approaches, it needs to be investigated in more detail.

4.3 Regression Analysis

We addressed the task of personality impression prediction with regression tasks targeted to predict the score of each one of the personality traits. We used support vector regression and followed a cross-validation approach by dividing the 281 samples in 10 folds, and using, at each resampling iteration, one fold for testing and the other 9 folds for training. Each time a model was trained, the parameters were optimized on the basis of another inner 10-fold cross validation.

We evaluated several models with distinct feature sets and different kernels. Results in Table 3 show R-squared prediction performance for experiments with a radial kernel (which provided only slightly better performance than other kernels). The three first rows of the table are experiments including all the cues, while the three last rows are experiments using significant cues from Table 1.

Not surprisingly, the best results are obtained for the Extraversion trait followed by Openness to Experience, which are the traits with large cue utilization. We observed that by using significant cues from the THR segmentation (THR_Sel) we obtained better performances than using significant cues from the HMM (HMM_Sel). In addition, the results suggest that combining cues from both segmentations may also help to improve results, at least for Extraversion. For the rest, the performance of HMMs and THR segmentations is comparable. Thus, we note that whereas HMM may be useful to interpret the presence and absence of facial expressions (see Section 4.1), it does not provide any advantage in terms of predictive performance, compared to using cues obtained from the THR segmentation.

4.4 Facial expressions vs. other cues

We were interested to assess the performance of facial expression activity cues to other cues used in previous work, and therefore, for comparative results, we have replicated regression experiments using the set of nonverbal features reported in [4]. Table 4 shows the result of predicting personality for our dataset, using three different sets of cues. As a first result, we found that the facial expression activity cues (THR_Sel+HMM_Sel) provided better performance than the visual activity cues (visual) used in [4], which estimate the total amount of the visual activity in vlogs. The result is relevant because both sets of cues extract information from the visual channel. However, the facial-based performance seems to be far from the one obtained from the audio, specially using prosodic cues.

5. CONCLUSIONS

We presented what to our knowledge is the first attempt to use fully automatic facial expression for the prediction of personality traits in vlogs. We used a state-of-the art automatic facial expression recognizer to process a sample of vlogs collected from YouTube and we modeled the presence of facial expressions using both a threshold and HMM-based methods to compute facial activity cues.

Overall, the results indicate that some facial expressions of emotion are indeed related to personality judgements, and that they are useful to predict impressions of Extraversion with better performance than using previously reported cues from visual activity and pose. We also found that, whereas the HMM segmentation seems to provide a more realistic characterization of the activation of facial expressions result in comparable, slightly lower performances than the threshold-based approach.

Acknowledgments This research was funded by the SNSF NCCR IM2 and Spanish Ministry of Education. We thank the Machine Perception Lab at UCSD for sharing CERT.

6. REFERENCES

[1] E. Agulla, E. Rua, J. Castro, D. Jimenez, and L. Rifon. Multimodal biometrics-based student attendance measurement in learning management systems. In *Proc. IEEE ISM*, 2009.

[2] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: Automatic assessment using short self-presentations. In *Proc. ICMI-MLMI*, 2011.

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

[4] J.-I. Biel and D. Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Trans. on Multimedia*, 2012.

[5] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Trans. on PAMI*, 21(10):974 –989, 1999.

[6] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259 – 275, 2003.

[7] S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big-five personality domains. *J. of Research in Personality*, 37:504–528, 2003.

[8] J. A. Hall, S. D. Gunnery, and S. A. Andrzejewski. Nonverbal emotion displays, communication modality, and the judgment of personality. *J. of Research on Personality*, 45(1):77 – 83, 2011.

[9] M. L. Knapp and J. Hall. *Nonverbal communication in human interaction*. Holt, Rinehart and Winston, New York, 2005.

[10] B. Knutson. Facial expressions of emotion influence interpersonal trait inferences. *J. of Nonverbal Behavior*, 20(3):165 – 182, 1996.

[11] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro. Modeling the personality of participants during group interactions. In *Proc. UMAP*, 2009.

[12] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In *Proc. IEEE FG*, 2011.

[13] D. McDuff, R. el Kaliouby, and R. Picard. Crowdsourced data collection of facial responses. *Proc. ICMI*, 2011.