What Your Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in YouTube

Lucía Teijeiro-Mosquera, Joan-Isaac Biel, José Luis Alba-Castro, and Daniel Gatica-Perez, *Member, IEEE*

Abstract—Social video sites where people share their opinions and feelings are increasing in popularity. The face is known to reveal important aspects of human psychological traits, so the understanding of how facial expressions relate to personal constructs is a relevant problem in social media. We present a study of the connections between automatically extracted facial expressions of emotion and impressions of Big-Five personality traits in YouTube vlogs (i.e., video blogs). We use the Computer Expression Recognition Toolbox (CERT) system to characterize users of conversational vlogs. From CERT temporal signals corresponding to instantaneously recognized facial expression categories, we propose and derive four sets of behavioral cues that characterize face statistics and dynamics in a compact way. The cue sets are first used in a correlation analysis to assess the relevance of each facial expression of emotion with respect to Big-Five impressions obtained from crowd-observers watching vlogs, and also as features for automatic personality impression prediction. Using a dataset of 281 vloggers, the study shows that while multiple facial expression cues have significant correlation with several of the Big-Five traits, they are only able to significantly predict Extraversion impressions with moderate values of R^2.

Index Terms—Face processing, facial expressions, personality prediction, vlogs

1 Introduction

The amount of multimedia data shared online everyday has exponentially increased in the last years. YouTube is one of the most successful examples, receiving 100 h of video every minute. The phenomenon of people uploading videos and other people watching them has created new types of social interaction. Conversational vlogging (video blogging) is a video genre where people record their opinions and feelings in a video and share this content with an audience.

In this article, we deal with the facial expression information shared in vlogging. Previous works have addressed the study of other nonverbal behavioural sources in vlogging including audio, gaze, and body cues [7], [8]. Facial expressions are a fundamental component in social interaction [46]. Humans use facial expressions to communicate their emotions, and to smooth or emphasize their points of view. Facial expressions are also commonly used to regulate communication [42].

The human face has been widely documented in the social psychology literature as an important source of information in interpersonal impressions [26], [29], [30]. By impressions, we mean the judgments that others make about a given person, in contrast to self-judgments. People rely on facial cues to make interpersonal judgments because there is a general belief that they convey valuable information about a person's character or personality [29]. In this paper, we examine personality impressions under the Big-Five model, that posits that human personality can be represented with five dimensions, namely extraversion (E), conscientiousness (C), openness to experience (O), agreeableness (A), and emotional stability (ES). The importance of the face information is especially true in the vlogging scenario, as vloggers typically show their head and shoulders on camera, and their faces occupy a large portion of the screen [8]. Among facial features, there is evidence that facial expressions of emotion provide information other than emotional states, influencing interpersonal impressions such as personality judgments, and that specific affective cues are in fact correlated with the possession of various personality traits [26], [30].

In the conversational vlogging setting, we present a systematic analysis of the capacity of facial expression cues extracted automatically with a state-of-the-art computer vision system to predict impressions of the Big-Five traits collected from external observers. A preliminary study was presented in [9], where we studied the prediction power of facial expressions using two basic types of facial expression cues extracted with the academically-available Computer Expression Recognition Toolbox (CERT) [33]. In this work, we perform a thorough analysis of the facial expressions of emotion in the vlog scenario; we assess CERT's performance using manually labeled data using crowdsourcing; we describe the cue content of vlogs and the relationship between facial expression cues and Big-Five personality impressions; and we perform regression experiments to automatically predict personality impressions from facial expression cues. The contributions of this paper are as follows:

- *L. Teijeiro-Mosquera and J.L. Alba-Castro are with Multimedia Technology Group, Vigo University, Pontevedra 36208, Spain. E-mail: {luciatm, jalba}@gts.uvigo.es.*
- *J. I. Biel and D. Gatica are with Idiap Research Institute, Martigny and EPFL, Lausanne, Switzerland. E-mail: {joan-isaac.biel, gatica}@idiap.ch.*

Manuscript received 26 Jan. 2014; revised 5 Aug. 2014; accepted 6 Oct. 2014. Date of publication 4 Dec. 2014; date of current version 3 June 2015. Recommended for acceptance by Q. Ji. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TAFFC.2014.2370044

- We propose and automatically extract four types of cues to predict personality impressions from facial expression. These cues characterize statistics of CERT outputs as dynamic signals over brief observation windows, inspired by existing literature on first impressions. We also study how to fuse the cue sets to improve the results through their combination.
- We systematically analyze the relationship (using correlation analysis) of each facial expression of emotion with impressions of the Big-Five traits obtained via crowdsourcing, concluding that Extraversion is the personality impression with more significant correlations regardless the cue extraction method.
- We study the prediction capability of the facial expression cues depending on their duration and relative location in the vlogs. We show that the time slices at the beginning of the videos predict better the annotators' impressions.

The rest of the manuscript is organized as follows. In Section 2, we survey the related literature. Section 3 outlines our approach, including the used dataset, the processing scheme, the cue extraction methods, and the fusion procedure. In Section 4, we present our experiments and results. Finally, we draw conclusions in Section 5.

2 Related Work

We review previous work related to this paper along three dimensions. First, we briefly discuss the state of automatic facial expression recognition. Second, we introduce the model used to characterize personality and we review recent work investigating the automatic recognition of personality impressions in vlogs. Third, we discuss work that has analyzed the effect of facial expressions of emotion on personality impressions.

2.1 Facial Expression Recognition

Facial expressions of emotion have been studied for decades. Already in 1872, Darwin [15] proposed that there are seven universal facial expressions of emotion that are produced and recognized for people all over the world, even for people from different cultures, sex, and races. The muscles involved in these facial expressions were studied by Duchenne using electrotherapy [17]. Since Ekman's formalization of the Facial Action Coding System [19], this method has become a standard framework in the computer vision community to code facial expressions of emotion. Using FACs, the movements of face muscles involved in expressions are coded in Action Units (AUs). The main advantage of using FACs is to provide scientists with a psychometrically validated tool to measure facial actions.

Research on facial expressions of emotion [18], [20], [45] in conjunction with advances in computer vision make it possible to develop tools that automatically recognize facial expressions. These tools are typically composed of a registration step followed by feature extraction and classification steps. The registration step detects the face and its relevant areas (forehead, eyes, mouth,...). There are two main approaches in registration: dense registration methods [34], [12], and coarse registration methods [33], [44]. The main difference between these two trends is that while dense registration methods invest a high effort in finding key points that allow to fully register the face, coarse registration methods rely on the invariance of textures descriptors to misalignment. Using dense registration, both shape [2] and texture [34], [12] could be used as features, but their performance critically depends on the registration step. Meanwhile, coarse registration methods are more robust to variation in lighting, strong movement of the face, and low video quality. Online data like vlogging is a very challenging scenario from the computer vision point of view [38]: videos are recorded in many different scenarios, with different points of view and varying illumination conditions. Because of this variability, we think that coarse registration methods are more suitable to process vlogs.

Given the above, for our work we decided to use the Computer Expression Recognition Toolbox, which detects facial expressions of emotion through texture-based AU classification [33]. CERT is composed of registration, facial feature extraction, and classification. CERT's registration phase consists of a face detector and ten facial landmark detectors. Using the information of the facial landmarks detected, the face is registered to a 96×96 grid using an affine warp. The facial features extracted for AU classification comprise 72 complex-valued Gabor filters with eight orientation and nine spatial frequencies. These features are classified into AUs using an SVM-based approach. The classifier provides the distance to the hyperplane that separates the two classes (AU_i activated or deactivated). Finally, facial expression of emotion recognition is done by feeding a multivariate logistic regression (MLR) classifier with the scoring from the AU classification. CERT produces outputs for all seven universal expressions (Fear, Disgust, Anger, Contempt, Joy, Surprise, and Sad) plus a Neutral expression. Moreover, CERT provides also a Smile detector based on boosting classification of haar-like features.

CERT has been applied in different areas of research, including discrimination of fake and real pain [5], detection of driver drowsiness [49], development of facial expression skills for autistic children [13], and understanding facial expressions during problem solving tasks [32]. Originally shared as an academic software package, CERT has recently become a commercial product, called FACET, which is enabling further studies.

2.2 Analyzing Personality Impressions

In this subsection, we first introduce the Big-Five Model and its role in the personality perception field. Afterwards, we describe previous work on the analysis of personality impressions in vlogs.

The Big-Five framework is a widely used model to characterize personality. This model organizes personality traits in five independent dimensions: Extraversion, Agreeableness, Conscientiousness, Neuroticism (N), and Openness to Experience [37]. In this work, we use Emotional Stability instead of Neuroticism to invert the scale and make all the Big-Five positive traits.

The two views of the personality perception field, personality impressions and self-reported personality, has been studied using the Big-Five Model [24]. Personality impressions explain how people see other people, while self-reported personality explains how people see themselves.

Previous research shows substantial convergence in some cases with self-reported personality, even if the impressions are formed from little information [11], [14]. This work focuses on personality impressions from short slices of social media data.

Regarding the analysis of personality impressions in vlogs, the current work extends recent research on the automatic analysis of nonverbal behaviour and personality impressions in vlogging [8], and contributes to a larger area of interpersonal perception research in social media [23] and social computing [31].

Previous research focused on collecting personality impressions from vloggers, automatically extracting nonverbal behavioural cues from audio and video, and automatically predicting personality impressions [8]. Regarding judgments of personality made by annotators, it was found that amongst the Big-Five traits, Extraversion and Agreeableness were the ones judged with highest accuracy in vlogging [7]. However, in the task of automatic personality impression prediction, nonverbal cues from audio and visual activity patterns seemed useful mainly to predict Extraversion (with R^2 values of up to $= 36\%$). These cues showed low performance for Openness to Experience and Conscientiousness ($R^2 = 10\%$ in both cases), and could not predict the Agreeableness trait. On the other hand, verbal content [10] showed power to predict Agreeableness ($R^2 = 31\%$), Conscientiousness ($R^2 = 19\%$), and Openness to Experience ($R^2 = 17\%$). Compared to past attempts to predict personality from audiovisual behaviour in meetings [31] and monologue presentations [6], the results in [8] suggested that the cues conveying personality information and the specific traits that can be reliably estimated using automatic analysis are particular to each communication scenario.

In contrast to the above works, the work presented here focuses on the extraction of cues of facial expression of emotion displayed by vloggers as a source of personal information, that to our knowledge has not been previously studied in the vlogging setting. While recent work has started to study online video, it has been either in the passive viewer case as in [38] (that analyzed observers of video advertising who essentially do not talk), or has used limited facial expression cues (smiles only) in the context of online video reviews (not addressing the personality inference task) [52], [40]. In contrast to these works, our work studies a much richer set of facial expression cues derived from all the basic facial expressions as estimated by a FACs-based recognizer. These facial expression-derived cues expand and complement the kind of audiovisual nonverbal cues investigated in all previous work. As stated in the introduction, a preliminary version of our study appeared as a short paper in [9]. In this paper, we further study this topic proposing two new cue sets that outperform the previous ones.

2.3 Personality Impressions and Facial Expressions of Emotion

Nonverbal behaviour research has investigated the many ways in which people use facial cues to make interpersonal impressions from others, through both static facial features (i.e. appearance) and dynamic facial expressions [29]. Studies have shown that amongst dynamic cues, people rely on the expression of emotion [4], [26], [28], [30], [39], [48], because there is a generalized understanding that these expressions not only provide information about people's affective states, but also convey information about personal traits. For example, a person who expresses happiness may be seen as someone who is confident, assertive, and friendly, whereas someone who is angry could be seen as an aggressive person. Nevertheless, research has also shown that impressions made on the basis of these facial expressions do not always agree with self-reported traits [26].

Earlier research investigated the influence of facial expressions on the basis of the Wiggins model, a framework that organizes personality using two orthogonal dimensions: dominance and affiliation [30], [39]. This model has connections to the Big-Five traits: Extraversion (resp. Introversion) correspond to high (resp. low) dominance and high (resp. low) affiliation, whereas the Agreeableness trait corresponds to the affiliation dimension [50]. These works concluded that people posing as happy and surprised are seen high in dominance and affiliation, whereas people showing anger and disgust are seen as high in dominance and low in affiliation. Other early research also investigated the links between smile and personality impressions, and has shown that people displaying smiles of enjoyment are judged as extraverted, emotionally stable, agreeable, sociable, pleasant, likable, and intelligent [11], [22], [36], [41].

More recently, Hall et al. [26] investigated how facial expressions influence attribution of the Big-Five personality trait impressions in three different conditions: people watching a video, narrating, and posing. Though the work aimed to identify cases where facial expressions of emotion are not diagnostic of self-reported personality (as personality impressions can differ from self-reported scores), the results show the importance that facial expressions play when making impressions from others.

Several of the works above have two main limitations. First, they focused on the study of facial expressions of emotion posed by actors [26], [30], [39], which raises questions regarding the strength of associations discernible in other settings. In comparison, we investigate a conversational setting that results from spontaneous video recordings and that is characterized by natural expressions. Second, these works approached facial expressions mostly as broad contextual conditions rather than measurable dynamic cues. To our knowledge, few works have investigated the effect of fine-grained facial expressions, mostly based on facial action units, on the personality impressions from computer animated characters [4] or other human trait inferences trait inferences such as dominance [28] or leadership [48] from posed expressions in photos. In contrast, our work contributes to the literature by analyzing the independent cue utilization of seven standard facial expressions of emotion. Furthermore, our work differs from all the above literature in that facial expressions of emotion are neither manually annotated nor posed, but automatically extracted using computer vision.

3 Experimental Setup

In this section we describe the vlog dataset and how this dataset is processed to extract sets of features. The vlogs are first processed using CERT to obtain temporal signals that

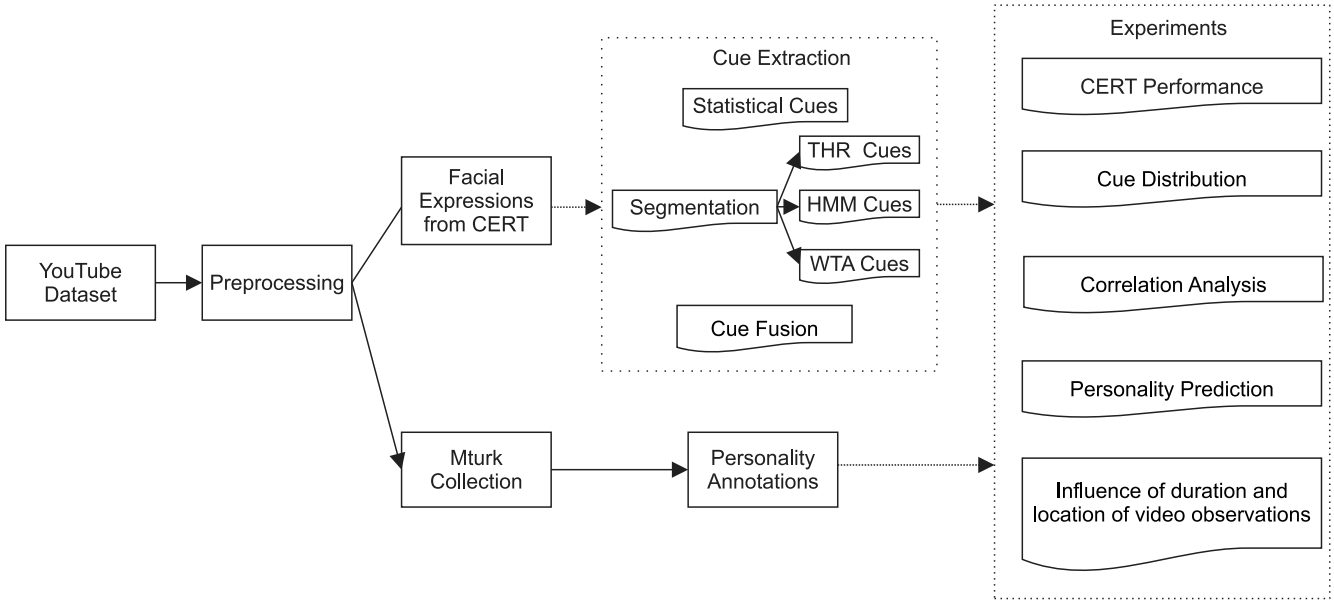

Fig. 1. Overview of our approach for the study of the influence of facial expressions of emotion in personality impressions.

indicate the frame-by-frame scores of each facial expression. Then, the signals are post-processed to extract condensed information that predict the personality impressions. The post-processing methods will be divided into two main groups: statistics-based and segmentation-based. The methods based on a previous segmentation of the CERT signal are further divided into three methods. All these methods produce what we called cues, so from this section and the rest of the paper we will be referring to cue extraction, cue distribution, cue selection, and cue fusion. Fig. 1 shows a summary of our approach to study the influence of facial expressions of emotion in personality impressions. In the rest of this section we describe the dataset, vlog processing, and feature extraction.

3.1 Dataset

The vlog dataset is composed of 442 videos from the same number of YouTube vloggers, and a collection of vlogger personality impressions, and was previously used in [8]. The videos feature a monologue scenario in which vloggers talk in front of the camera during one minute, mainly showing head and shoulders, and display spontaneous behaviour. The videos have different framerates from 6 to 30 fps. All the videos were clipped to one minute duration. The dataset is balanced in gender, with 208 males (47 percent) and 234 females (53 percent). Though the vlog dataset has 442 videos, only the 281 videos with better registration performance are used in our experiments.

The personality impressions were collected in [8] using Amazon Mechanical Turk. The annotation task consisted on watching a vlog and, after finishing, answering a short personality questionnaire designed to measure impressions of the Big-Five personality traits: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience. For each vlog, personality impressions were aggregated from five annotators and show the following intraclass correlation, ICC(1,k): .76 (E), .63 (A) .42 (C), .40 (ES), and .49 (O). It is worth mentioning that reliabilities compare well to those found in other personality impression settings [8].

3.2 Vlog Processing

The realistic conditions of vlogs make it a very challenging scenario for computer vision. We use CERT, that is based on coarse registration methods, to recognize facial expressions. The abundance of recording scenarios, the multiple camera positions, and the continuous illumination changes suggest the use of methods robust to misalignment.

Even so, in order to minimize problems during the face registration step of CERT, we selected those videos with a better registration performance. First, we used the Viola-Jones face and facial feature detector with very high precision (low rate of false positives) to detect face, eyes, nose, and mouth [1]. Second, we picked those videos where all the five facial features were located in a high percentage of frames. In the selected subset, the average rate of frames with all the features detected was 75 percent with an standard deviation of 19 percent. The video with the worst percentage of detections had the five features detected in the 25 percent of their frames, while the video with the best performance had the five features detected the 99 percent of the time. This pre-processing step aimed to minimize the error in facial expression recognition results due to poor registration. We tested if there were any differences among the registration categories regarding personality impressions, e.g. whether videos with poor registration might be associated to people with low rate of conscientiousness. However, no significant effects were found.

The final subset contains 281 vloggers, balanced in gender, and with similar reliabilities for the personality impressions compared to the complete database. The 281 videos were processed using CERT. The outputs of CERT are time series with frame-by-frame estimates for each of the facial expressions. The facial expression scores are normalized between zero (no facial expression) and one (maximum intensity). On the other hand, the Smile signal provided by

CERT varies between negative values (no smile) and positive values (smile).

We process the CERT signals to extract a set of higher-level facial expression of emotion cues for every video sequence.

3.3 Cue Extraction

In this section we propose four different methods to aggregate the CERT temporal signals into a set of cues that characterize the amount and activity patterns of facial expression of emotion for each video.

The first method computes basic statistics of the CERT signal, a representation that characterizes the distribution of the CERT values but that is time independent. The three other methods aim to quantify the presence or absence of the facial expression of emotion along the video sequence. These cues are related to the quantity of time the facial expression is active and the number of occurrences of the facial expression.

3.3.1 Statistic-Based Cues

The statistical cues consist of calculating seven statistic values over each CERT signal. The seven statistics from each facial expression are: Mean, $\mathrm{Variance}$, Median, $\mathrm{Maximum}$, $\mathrm{Minimum}$, $\mathrm{Entropy}$ and $\frac{\mathrm{Var}}{\mathrm{Mean}}$. This approach provides a set of 63 cues, given by the combination of each statistic and each facial expression of emotion (Mean_{Anger}, Var_{Sad}, Mean_{Joy},...). Using these cues, we try to represent the facial expression patterns of each video. Mean, Median, Maximum and Minimum characterize the signal amplitude, while variance, entropy and $\frac{\mathrm{Var}}{\mathrm{Mean}}$, represent the signal variation over time.

3.3.2 Activity Cues

The activity cues are calculated over the segmented CERT signal, i.e. a binarization of CERT's output that indicates whether the facial expression is active or not at the given frame. In this work, we propose three different segmentation methods: Threshold (THR) cues, Hidden Markov Models (HMM) cues, and Winner Takes All (WTA) cues.

Threshold Seg. Thresholding the signal is the straightforward solution to decide the activation state from CERT output for each facial expression. Threshold segmentation tracks high-frequency changes on the facial expression signal, but it is sensitive to outliers, i.e., frames with poor registration, and hence poor facial expression recognition that produces a noisy segmentation. In our approach, we set the threshold to a low value so that low signal values are taken as inactive. The threshold is low in order to keep most of the active segments. Meanwhile, for the smile detector, we use a zero threshold imposed by design (see Section 2.1).

HMM Seg. We use a two-state HMMs to detect the active an inactive state for each CERT output. Each state is modeled with one Gaussian initialized with the threshold-based segmentation, while the transition probabilities are set to $\rho_{00} = \rho 11 = .95$ and $\rho_{01} = \rho 10 = .05$. In practice, the THR approach copes with high frequency changes, tends to generate 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.

TABLE 1
Summary of Cue Sets

Cue Set	Cue	Number of cues
Stat. Cues	Mean, Var, Median, Max, Min, $\mathrm{Entropy}$, $\frac{\mathrm{Var}}{\mathrm{Mean}}$: Mean_{Smile}, Var_{Joy}, Median_{Sad}, ...	7x9 = 63
THR Cues	PT, AD, PTS, NS : PT_{Joy}, AD_{Anger}, ...	4x9 = 36
HMM Cues	PT, AD, PTS, NS : NS_{Sad}, $\mathrm{PTS}_{Neutral}$, ...	4x9 = 36
WTA Cues	PT, AD, PTS, NS : PT_{Smile}, AD_{Surp}, ...	4x9 = 36

WTA Seg. Finally, the WTA segmentation is designed to avoid concurrence of the facial expression activation. Using HMM and THR, facial expressions are processed independently, so two or more facial expressions can be active at the same time. WTA is an alternative segmentation that only keeps active the signal with the highest score and makes all the rest inactive.

Cues. Let r be the state of the segmented signal ($r = 1$, active; $r = 0$, inactive), where one segment is the collection of consecutive frames with the same state. We define:

- Proportion of active time (PT): computed as

$$PT = \frac{1}{N}\sum_{i=1}^{N_r} \tau(r_i = 1), \tag{1}$$

  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.
- Rate of active segments (NS): computed as

$$NS = \frac{f}{N}\sum_{i=1}^{N_r} (r_i = 1), \tag{2}$$

  where f is the frame rate.
- Average duration of active segments (AD): computed as

$$AD = \frac{1}{N_r}\sum_{i=1}^{N_r} \tau(r_i = 1). \tag{3}$$

- Proportion of time with short active segments (PTS): computed as

$$PTS = \frac{1}{N}\sum_{i=1}^{N_r} \tau(r_i = 1 | \tau(r_i) \leq .1f), \tag{4}$$

  where $0.1f$ corresponds to 100 ms; i.e., the proportion of time in segments shorter than 100 ms.

In summary, the activity cues or segmentation-based cues measure: the percentage of time each facial expression of emotion is active (PT), the frequency an active segment appears in the facial expression signal (NS), the average duration of the facial expression segments (AD) and the percentage of time the signal is active in segments shorter than 100 ms (PTS). These four cues are extracted for each facial expression of emotion making up, a whole cue set of 36 cues (PT_{Anger}, PT_{Smile},...) per segmentation type. Table 1 summarizes the cue sets.

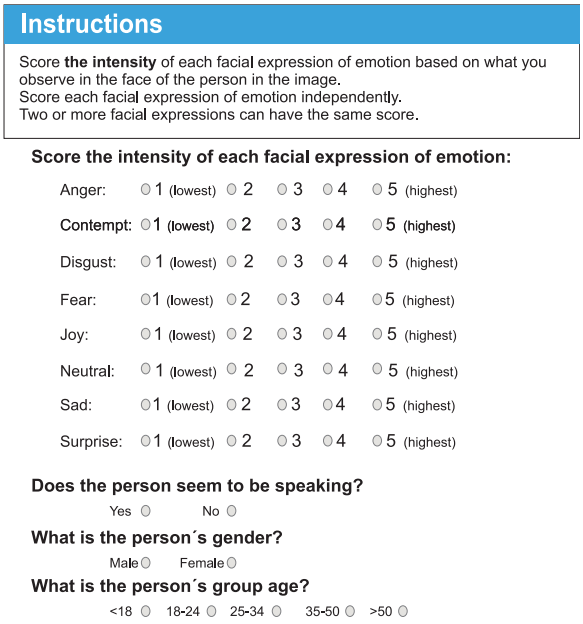

Fig. 2. Form filled by MTurk workers while looking at the vlogger's image.

3.3.3 Fusion of Cues

Together with the feature sets above, we investigated the fusion of features to combine the potentially different information captured by them. In principle, we would expect better results if this information is complementary. We study the following approaches:

- We compare different combination of cues: THR +HMM, THR+HHM+WTA, Statistical+THR+HMM +WTA, and Statistical+WTA.
- We calculate a new cue set: Statistics of Active Time. In this cue set, we calculate both the activity cues and the statistical cues over the segmented facial expression signal.

4 Experiments and Results

We now present the experiments and results of our study. The section guides the reader from the analysis of the raw CERT signals to the prediction of personality impressions from the facial expression cues. In the first subsection, we analyze CERT's output to understand if the information it conveys is coherent enough for our study. In the second and third subsections, we study how the cue extraction methods represent the facial expression signals, and their correlation with the personality annotations. In the rest of the subsections, we address the task of predicting personality and evaluate the influence of the facial expression's time slices according to their duration and relative location in the vlogs.

4.1 CERT Assessment in Vlogs

We want to analyze if CERT's reliability in the vlogging scenario is enough to carry out this study. Although such reliability has been demonstrated in [33] to recognize posed facial expressions of emotion in Cohn-Kanade dataset [27] and spontaneous action units on M3 database [21], the vlogging scenario is particularly difficult for registration and thus, for facial expression recognition. Given that the large amount of data prevents us from doing frame by frame annotation, we selected a subset of frames that were annotated through crowdsourcing. We selected frames from seven categories: Anger, Contempt, Disgust, Fear, Joy, Sad and Surprise. These frames were selected using CERT scores as representative frames for each facial expression of emotion. For each category, we selected 200 frames that fulfilled the following criteria:

- The score for its category was the highest among all the facial expressions of emotion.
- Only frames with scores over the third quartile were selected.
- To increase variability, no frames separated less than 100 ms were allowed .
- In order to have at least 60 different vlogs represented in the subset, the number of frames per subject was upper-bounded.

The images were annotated using Amazon Mechanical Turk. In the experiment, MTurk workers were asked to look at the image and score the intensity of each facial expression of emotion independently. The Human Intelligence Task (HIT) was designed to show one image and the questionnaire at the same time. Fig. 2 shows the questionnaire but not the vlogger image due to data protection issues. Two control questions about demographics (age group and gender) were added to add control to the data with respect to spammers. Each image was annotated by 5 different workers, with a reward of $0.05 per image annotated. The full set of 7,000 annotations was collected in about 72 hours, with the participation of 73 different workers. The average annotation time was 39.4 s. The HITs were restricted to US workers with HIT acceptance rates of 95 percent or higher. It is worth to mention that workers were not trained in facial expressions of emotion and therefore, results are constrained to the normal ability of people to identify these face expressions.

TABLE 2
Intraclass Correlation for Facial Expression Scores

	Anger	Cont.	Disgust	Fear	Joy	Sad	Surp.
All Images (1400)	.66	.48	.71	.53	.90	.59	.70
Only Images Same FE (200)	.80	.35	.89	.70	.81	.61	.74

In the first row the ICC(1,k) was calculated over all the images, while in the second row only the frames of the facial expression of emotion selected by CERT were used.

Table 2 shows the intraclass correlation coefficient [47] of the annotated data. In the first row, we show the ICC(1,k) calculated over the whole dataset (1,400 images), while in the second row the ICC(1,k) is calculated only over the representative images for each facial expression of emotion (200 images). In the whole dataset, Joy shows the highest ICC reliability followed by Disgust and Surprise, which indicates that observers agree more when annotating these facial expressions. Comparing the first and second rows, we observe that for all the expressions except for Joy and Contempt the ICC is higher when calculating the ICC only in the images that, according to CERT, show that facial

TABLE 3
Confusion Matrix between MTurk Annotations and CERT Categories

MTurk \ CERT	Anger	Cont.	Disgust	Fear	Joy	Sad	Surp.
Anger	**21.0**	0.5	6.0	4.0	1.0	5.0	3.5
Cont.	7.0	**11.0**	2.0	3.0	1.5	11.0	1.0
Disgust	5.5	1.0	**22.5**	4.0	4.0	5.0	2.5
Fear	1.5	1.0	0.0	**5.0**	1.0	4.0	5.0
Joy	5.0	31.0	11.0	18.0	**83.5**	12.5	10.0
Sad	3.5	1.5	1.0	2.0	0.5	**22.5**	0.5
Surp.	1.0	4.0	1.0	21.5	1.5	5.5	**33.0**
Neutral	55.50	50.0	56.5	42.5	7.0	34.5	44.5

Each column corresponds to the facial expression selected using CERT, and each row corresponds to the selection made by annotators.

expression of emotion. This increment indicates that annotations are less noisy when the workers are exposed to the most likely facial expression they are annotating, and therefore that CERT is performing a fair selection. Moreover, the low ICC values of the first row point that, except for Joy, scoring spontaneous facial expression of emotion of a single image is not a trivial task. Finally, the ICC value of Contempt indicates that this expression is the hardest one recognize in general but also in the subset of frames selected using CERT.

CERT has been showed to have a 76.1 percent recognition performance over seven facial expressions of emotion from the Cohn-Kanade database [33]. The best classification rate was for Joy, Disgust, and Surprise, while the worst classification rate was for Anger (Contempt was not studied). These results do not need to generalize to the vlog dataset, as both datasets are different in nature. We performed a similar study in our dataset. The scores for each facial expression of emotion are aggregated using the mean of the five workers' annotations. The facial expression with the highest score is chosen as the winner to calculate the classification results.

Table 3 shows the confusion matrix of the facial expressions. Each column corresponds to the facial expressions categorized using CERT, while each row corresponds to the MTurk workers' decision. The performance is lower than in the Cohn-Kanade dataset, with an average success of 28.3 percent. However, all the facial expressions but Fear and Contempt are classified over chance $(1/8 = 12.5\%)$ using untrained annotators. Regarding the misclassified frames, Neutral expression appears as the winner for all the facial expressions, except for Joy that, as the ICC results pointed out, is the facial expression of emotion with the best performance. The fact that a significant number of the frames were considered Neutral reflects that the annotators could not find strong evidence to decide on the facial expression selected by CERT and they scored it, on average, with lower intensity than neutral. This is in accordance with the human lower performance in classifying spontaneous facial expressions of emotion [25], like those appearing in the Vlog scenario. On the other hand, Fear is the facial expression of emotion with the worst classification rate. The results show that, as in Cohn-Kanade dataset, a high percentage of the Fear frames were confused with Surprise. Besides the low performance of CERT classifying Fear, the misclassification between Fear and Surprise could be also affected by the difficulty of humans to distinguish between these two emotions [43]. Finally, concerning Contempt, which was not analyzed with CERT over Cohn-Kanade in [33], on the vlog data it also shows a low classification rate, being confused mostly with Joy. The poor performance of Contempt might be related to the difficulty in the detection of correct mouth-related AUs caused by the talking scenario.

In conclusion, the results show that CERT's facial expression recognition performance is acceptable to carry out our study, and that the facial expression recognition problem in unconstrained online social video is a challenging issue. All facial expressions of emotion, except Fear and Contempt, showed a classification rate above 20 percent. Moreover, Joy shows the best performance in CERT.

4.2 Distribution of Facial Expression Cues

In this section, we explore the distribution of facial expression cues in vlogs using the four methods proposed. This study is useful to identify the amount of facial expression activity in our dataset.

Fig. 3 shows the histogram of the Statistical Cues. Neutral is the facial expression with higher means and, hence, the facial expression with more presence in the vlogs. Joy, Fear and Surprise are also present in the dataset, although with smaller means. The distributions of Anger and Disgust suggest that these are the least frequent expressions in vlogs. Sad and Contempt have higher means and more variance than Joy, Fear and Surprise, however we could not find evidence in the data to support the higher presence of these facial expressions of emotion. One of the most likely explanations is that the higher levels of Contempt are caused by

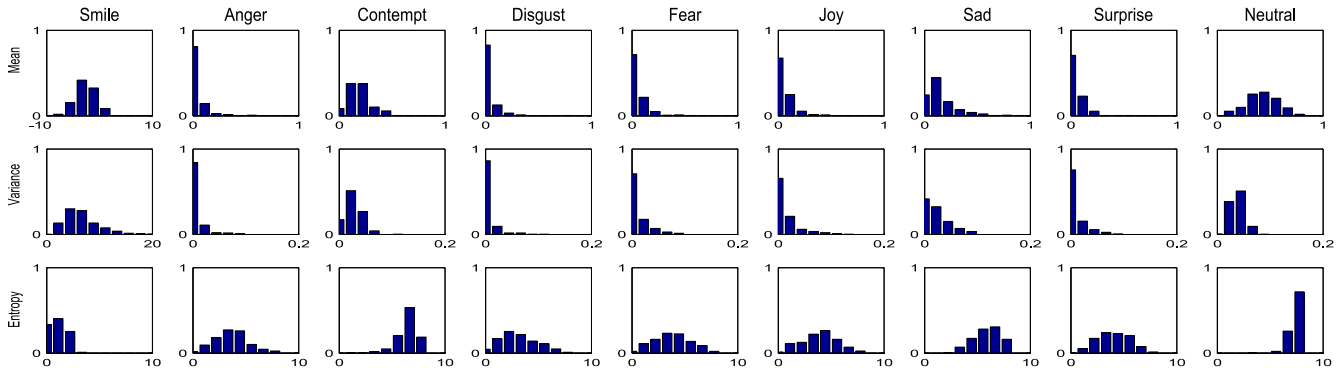

Fig. 3. Histogram of selected statistical cues. Each plot represents the distribution of one cue, where the x-axis shows the range of the cue, and the y-axis shows the ratio of vlogs with a given cue value.

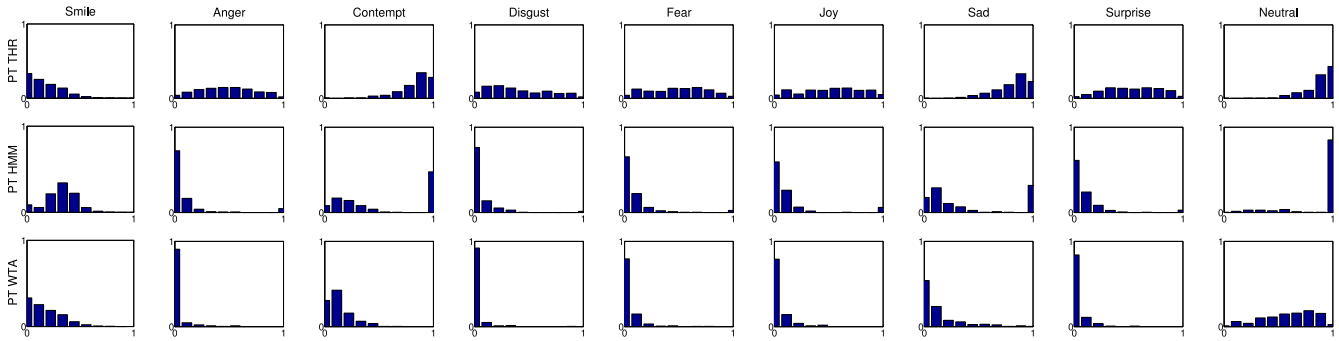

Fig. 4. Histograms of the PT cue for the different segmentation methods: THR, HMM, and WTA (from top to bottom). The *x* axis represent the range of the cue and the y axis represent the ratio of vlogs with a given cue value.

noisy facial expression recognition in the talking scenario, as shown in the previous section. In the case of Sad, we also hypothesize that the cue extraction method is overrating its presence in the dataset. Note that the Smile distribution has a different range due to the use of a different detector. Summing up, the distributions of the cues suggest that they convey enough variability (information) about the facial expressions to use them for further processing.

Fig. 4 shows the distribution of activity cues using Threshold, HMM and WTA. We focus on the analysis of the portion of active time (PT) for clarity reasons. First, we compare the distributions of THR and HMM cues. We observe that the PT distributions vary depending on the segmentation method. Anger, Fear, Disgust, Joy and Surprise, whose distributions are more spread using THR cues, are concentrated around small values using HMM cues. On the contrary, the distribution of Neutral, that is more spread using THR, is concentrated around one using HMM, an indication that Neutral is active almost the whole video. Again, Contempt and Sad show different distributions from the other facial expressions. In the case of THR segmentation, for example, the values of PT for Anger, Joy, Fear, and Surprise are spread between zero and one, while for Contempt and Sad, PT values are more concentrated around one.

Regarding WTA segmentation, we observe that the distributions of Anger, Disgust, Fear, Joy, and Surprise are quite similar to those corresponding to the HMM segmentation. Meanwhile, the distribution of Contempt, Sad, and Neutral differ from both the HMM segmentation and the THR segmentation. We suggest that WTA segmentation could help smooth the results of Sad caused by the talking scenario, representing better than the other cue extraction methods its presence in the dataset. It is interesting to note that Neutral is the facial expression with signal activity, although when segmented with WTA, it has less activity than using HMM or THR, indicating that there are many frames containing other winning expressions. We argue that the high presence of Neutral might be caused by the nature of vlogging, where people are most of the time looking with a relative Neutral expression to the camera.

All the facial expression categories are detected in the vlog dataset. In the case of Contempt, we believe that this expression, as shown in the previous section is poorly recognized in scenarios with talking faces, due to the low discriminative power of the upper face in this expression and the interference of mouth movement while talking. This also stands for Sad, which seems to be more sensitive to the cue extraction method. Regarding Neutral expression, we suggest that its high occurrence appropriately represents the vlog scenario, where people mainly have Neutral expression and the expressions segments are short.

We also explored the amount of overlap between segmented signals. We define the co-occurrence of expression e_i with expression e_j, as the percentage of time both expressions are active divided by the time the expression e_j is active. Formally, this is expressed as:

$$P(e_i|e_j) = \frac{\sum_{k=0}^{N_r} \tau(r_k^i = 1, r_k^j = 1)}{\sum_{k=0}^{N_r} \tau(r_k^j = 1)}, \tag{5}$$

where r^i is the state of the segmented signal i, N_r is the number of segments and $\tau(r)$ is the duration of the segment r. Table 4 shows the co-occurrence between each pair of facial expressions using the HMM segmentation. Note that this measure is not symmetric (for example, the occurrence of smile with surprise is .25 and the co-occurrence of surprise with smile is .07). The main diagonal shows the mean percentage of time each facial expression is active according to the HMM segmentation, i.e., it does not corresponds to $p(e_i|e_i)$. The co-occurrence results are presented using HMM segmentation instead of Threshold segmentation to help the analysis. As we commented before, the THR segmentation was designed to be highly permissive for facial expression activation. Hence, it produces higher levels of co-occurrence that are difficult to interpret.

TABLE 4
Co-Occurrent Facial Expressions

	Smile	Anger	Cont.	Disg.	Fear	Joy	Sad	Surp.	Neutral
Smile	.31	.08	.61	.06	.09	.22	.40	.07	.87
Anger	.27	.09	.50	.23	.08	.10	.48	.10	.86
Cont.	.33	.08	.58	.04	.04	.13	.37	.07	.97
Disgust	.35	.38	.39	.06	.09	.04	.47	.05	.80
Fear	.33	.08	.31	.06	.08	.16	.49	.17	.70
Joy	.59	.08	.63	.02	.12	.12	.38	.13	.79
Sad	.30	.10	.50	.06	.10	.10	.42	.09	.91
Surp.	.25	.11	.43	.03	.16	.16	.43	.09	.92
Neutral	.30	.09	.62	.05	.07	.10	.42	.09	.91

Columns represent the probability that each expression occurs given that the facial expression in each row is active. For example, $P(Anger|Smile) = .08$ and $P(Smile|Anger) = .27$. The diagonal represents the mean PT of the facial expression.

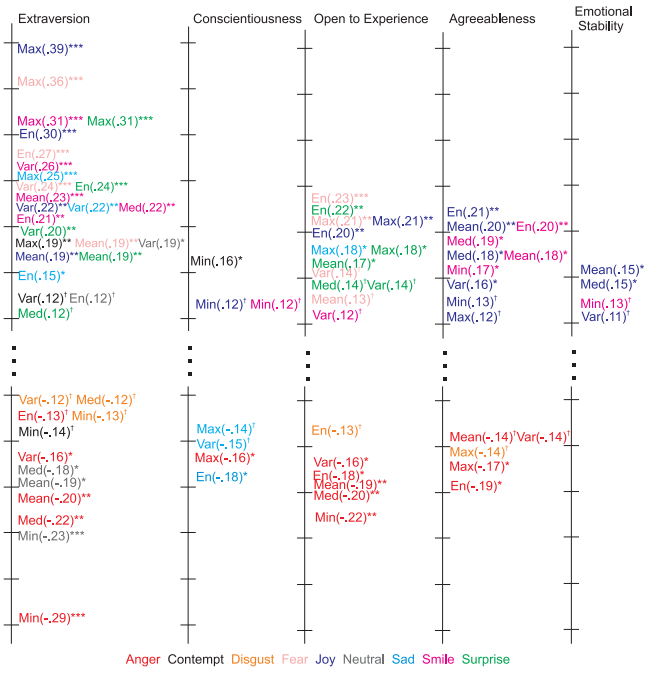

Fig. 5. Correlation effects between statistical facial expression cues and personality impressions. Correlation values are ordered, and each color represents one facial expression. ($^{\dagger}p < .05, ^{*}p < .01, ^{**}p < .001, ^{***}p < .0001$).

An interesting result that suggests a good segmentation process, is the relation between Joy and Smile—the co-occurrence of Joy is higher given Smile (.22) than given any other facial expression and also the co-occurrence of Smile is higher given Joy (.59) than any other facial expression. A high co-occurrence is also observed between Anger and Disgust, and between Fear and Surprise, although we could not find evidence that this co-occurrence is caused by the presence of compound facial expressions of emotion [16]. The co-occurrence between Fear and Surprise might be caused by the poor recognition of Fear shown in the previous section. On the contrary to THR and HMM, the third segmentation approach presented in Section 3.3 (WTA) does not allow, by design, any co-occurrence.

4.3 Correlation Analysis between Facial Expression of Emotions and Personality Impressions

In this section, we analyze the correlation between the facial expression cues and the personality impressions.

Only correlation coefficients with $\mathrm{p-value} < 0.05$ are used: we call them correlation effects. The correlation effects of the Statistical cues are shown in Fig. 5, where the correlation values are ordered from top-higher positive to bottom-higher negative, and a different color for each facial expression of emotion for readability.[1] As with previously used audiovisual cues [7], [8], the facial expressions of emotion showed higher correlation values for Extraversion independently of the representation. This trait was the one with the higher number of correlation effects (STATS, significant correlations = 37; THR, significant correlations = 18; HMM, significant correlations = 20; WTA, significant correlations = 21), followed by Openness to Experience and Agreeableness.

The Extraversion trait was mostly negatively correlated with cues that express Anger (*Mean*; −.20; thr-PT: −.16; hmm-PT: −.22; wta-PT: −.19), and Disgust (*Mean*: −.12; thr-PT; −.13; hmm-PT: −.12), and positively correlated with Joy (*Mean*: .19; *Max*: .39; thr-PT: .23; hmm-PT: .23; wta-PTS: .27) and Smile (*Mean*: .23; thr-PT: .25; hmm-PT: .23) which concurs with the idea that Extraverted people are more enthusiastic. However, other effects may be more difficult to explain, such as the positive correlation with Sad (*Max*: .25; hmm-PT: .25). Openness to Experience also showed similar negative correlations for Anger (*Mean*: −.19; thr-PT: −.20; hmm-PT: −.16), but showed only a couple of effects with Joy and Smile. We also observed positive correlations between Openness to Experience and Fear, which concurs with the effects in Extraversion and may suggest that this facial expression is not correctly estimated, as previously discussed.

The Agreeableness trait is also negatively correlated with Anger (*Mean*: −.14; PT: −.16) and positively correlated with Joy (*Mean*: .20; PT: .18), and Smile (*Mean*: .18; Entropy: .20; thr-PT: .20), and did not show any effects with any other expressions. Finally, Conscientiousness and Emotional Stability showed a very small number of effects. Overall, we found that THR and HMM features provided similar effects in terms of the sign, though the cue utilization value varied across facial expressions and traits.

In summary, whereas CERT seems to be capturing information that agrees with impressions of personality, how much the method suffers from processing challenging conversational social video like the one we study remains an open question. In particular, the fact that most vloggers talk during a substantial amount of time may trigger some facial expressions due to lip movements that would not be otherwise activated [35]. These issues ned to be investigated in future work.

4.4 Personality Prediction

We address the task of predicting personality impressions from facial expression cues. We used support vector regression to predict each personality impression independently. The SVM regressor is trained following a double cross-validation approach, by dividing the 281 vlog samples in 10 folds and using, at each resampling iteration, one fold for testing and the other nine folds for training. Each time a model was trained, the SVM parameters were optimized on the basis of another inner 10-fold cross validation.

We measured the performance of the system in terms of the coefficient of determination R^2. This coefficient is computed as the ratio between the model prediction and the model baseline ($\bar{y}_{obs}$). In other words R^2 expresses the quantity of variance explained by the model

$$R^2 = 100\left(1 - \frac{\sum(y_{obs} - y_{pred})^2}{\sum(y_{obs} - \bar{y}_{obs})^2}\right). \quad (6)$$

Note that $\bar{y}_{obs}$ corresponds to the mean over the training data, and not the whole data, as in preliminary results presented in [9]. This decision makes the regression results different than in [9] although it represents

1. The Table with the whole set of correlation effects for each cue extraction method is provided as additional material.

TABLE 5
Regression Comparative of Each Feature Set

	Extr	Cons	Open	Agr	Emot
Statistics	.15*	-.11	.07	-.04	-.10
THR	.13**	-.05	.03	-.01	-.23
HMM	.09*	-.22	.05	-.06	-.15
WTA	.17***	-.07	.04	-.04	-.15

$(^{*})\ p < .05$, $(^{**})\ p < .01$,$(^{***})p < .001$.

the testing scenario better. This also causes the occurrence of negative values in R^2.

We evaluated several models with distinct feature sets and different kernels. Results in Table 5 show the prediction performance for experiments with a radial kernel (which provided only slightly better performance than other kernels). For these experiments we used all the cues of each cue set. The $\mathrm{p-values}$ shown in Tables 5 and 6 are calculated using a two-tailed single t-tests to measure significant differences between the models and the baseline.

As shown in Table 5, only the Extraversion impression could be predicted with statistical significance and a moderate value of R^2. The results concur with previous attempts to predict personality using audiovisual features, and indicates that Extraversion is easier to judge using this type of behavioral information. Amongst all feature sets, we found that Statistical and WTA cues outperform THR and HMM cues. It is worth mentioning that the ability of the WTA segmentation to fit the dynamics of CERT signals may be beneficial for personality prediction. This specific issue needs to be further investigated in future work. The low prediction performance for the other personality impressions needs further investigation to understand if their performance is affected by their low agreement or because facial expressions of emotion are not appropriate features. Recent results presented in [10] suggest that verbal cues are more suitable to predict Agreeableness, Conscientiousness, and Open to Experience. Moreover, it would be interesting to test if the results are affected by the lower ICC reliability of the other personality traits compared to Extraversion. Future work could investigate differences between predicting each individual annotators' impression and predicting the aggregated impression.

Given that different cue sets have different distributions, we hypothesize that a combination of cue sets could improve the performance. Table 6 shows the results on regression performance for different combinations of cues. The first row corresponds to the combination of THR and HMM cues; this combination does not improve much the performance with respect to THR. The second row shows the combination of the three segmentation based methods and the third row shows the combination of the four cue sets. None of the combinations manage to outperform WTA-only, that is the best cue set for predicting personality.

TABLE 6
Regression Comparative of Feature Set Combinations

	Extr	Cons	Open	Agr	Emot
THR & HMM	.14**	-.10	.04	-.05	-.17
THR, HMM & WTA	.16***	-.10	.05	-.00	-.14
Stat., THR, HMM & WTA	.17***	-.07	.07	-.01	-.12
Stat. & WTA	.17**	-.15	.06	-.04	-.14
Stat. of active time	.19***	-.10	.07	-.05	-.16

$(^{*})\ p < .05$, $(^{**})\ p < .01$,$(^{***})\ p < .001$.

TABLE 7
Regression Results for Extraversion Dividing the One-Minute vlogs Into Shorter Time Slices

	Extraversion			
Two Segments	A_1	B_1		
	.17	.11		
Four Segments	A_2	B_2	C_2	D_2
	.14	.13	.13	.08

Finally, segmentation-based cues like WTA do not take into account the intensities of the facial expression, so a combination with Statistical cues could improve results. However, the combination of Statistic and WTA methods does not seem to improve the prediction. The last row of Table 6 shows the regression performance using only the Statistics of active time. This provides slightly better results than WTA, suggesting that the Statistics could be more informative when they are extracted from the segmented signal. However, the increase in performance is not statistically significant.

4.5 Influence of the Slice Duration and Location

Finally, we study the potential predictive power of facial expressions depending on the duration and relative position of the specific vlog segment under consideration (i.e., the amount of observations and their position), by replicating the prediction experiments for different vlog slices.

We perform two experiments. In the first one, we divide the one-minute vlogs into two slices of 30 seconds, referred to as $A1$ and $B1$ according to their position in the vlog. In the second experiment, we divide the vlogs into four slices of 15 seconds (A_2, B_2, C_2, and D_2). For each slice selection, we train and test an SVM regressor using the experimental procedure explained in the previous section.

Table 7 shows the R-squared prediction performance for the extraversion impression (for the other impressions, results are not significant). In our first experiment, we observe that better results are achieved using the A_1 slice than using the B_1 slice. In the second experiment, the performance of A_2, B_2, and C_2 slices are very similar, but drops substantially for the last slice D_2. These results show that viewers' impressions are better predicted by features computed at the beginning of each vlog. This result concurs with the idea that first impressions are built from short interactions [3], [51] and suggests that not much information might be used at the end of a vlog to build impressions. In the future, this result could potentially be used to limit the extent of automatic processing of vlogs without decreasing performance, which can be useful for computationally expensive feature extraction methods. Nevertheless, further research needs to be done to confirm this first result. For instance, it would be interesting to test if the same effect is observed for every nonverbal cue source (audio, facial or multimodal), and whether the optimal duration and position of the vlog slices are the same for each data type.

5 Final Discussion and Conclusions

In the context of social video analytics, we presented what to our knowledge is the first attempt to use fully automatic facial expression recognition for the prediction of personality trait impressions in conversational vlogs. We rely on a state-of-the art automatic facial expression recognizer to process a sample of vlogs collected from YouTube, and provided different methods to characterize the facial expression content of vlogs.

We first assessed CERT's performance in vlogs through the evaluation with manually annotated data. We found that Joy is the facial expression of emotion with the best performance. Besides, the experiment demonstrated that the facial expressions of emotion automatically detected in the vlogs, except for Fear and Contempt, are acceptable to be further processed.

We then characterized the facial expression content of vlogs using four cue extraction methods that reflect different statistical and temporal features of the CERT signals. Through this work, we have shown that facial expressions of emotion have significant correlation with personality impressions, specially with extraversion.

Furthermore, we demonstrated that extraversion impression can be predicted with $R^2 = .19$ using automatically extracted facial expressions of emotion cues. Extraversion is the best predicted trait regardless of the cue extraction method. We compared the four cue sets, and found that WTA cues outperform the other methods. We have shown how the high frequency component of facial expressions is important to predict extraversion. Moreover, we have shown how to improve WTA performance by the combination of Statistical and WTA cues. On the other hand, none of the other four traits of the Big-Five model could be predicted with the proposed cues. This is an issue that also arised in past work using other audio-visual behavioural cues [8]. This could be due to the fact that annotators do not rely on cues similar to the ones we extracted to judge the traits, and also due to errors in cue extraction. Interestingly, recent work has shown that the verbal content of vlogs can predict other traits rather than extraversion [10], which points towards the possibility to use both verbal and nonverbal features.

Finally, we studied the influence of the duration and relative location of the observed facial expressions. We showed that competitive prediction results for extraversion could be obtained with shorter time slices. Also, we showed that the slices at the beginning of the video predict better viewers' impressions. These results prompt interesting questions, for instance, if the same happens for other nonverbal cues. We suggest that this effect might be caused by the viewers' being less sensitive to the facial expressions after making up their first impressions. This issue needs to be studied in detail in future work.

Regarding future work, we acknowledge a main shortcoming in our study, which is the evaluation of the influence of the talking scenario. This problem could be investigated in future work by exploring the output of CERT on speech and non-speech segments using an automatic speech/non-speech detector or a finer representation of the verbal content [35]. As a second issue, in this work, personality impressions were treated as independent signals. However, it would be interesting to analyze the overall perception and see if the personality impression about one trait could influence the impressions about other traits. Moreover, the study was limited by having only one vlog per user. Finally, the superior performance of basic statistics compared to most of the segmentation-based approaches may also motivate further work on alternative statistical representations that exploit the distribution of features.

Acknowledgments

This work was partly conducted while the first author visited Idiap. The authors thank the support of the Swiss National Science Foundation through the National Center of Competence in Research on Interactive Multimodal Information Management (IM2) and Spanish Ministry of Education under the project CN 2012/260 "Consolidation of Research Units: AtlanTIC", and by the Spanish Ministry of Economy and Competitivity under the project TEC2012-38939-C03-01. L. Teijeiro Mosquera is the corresponding author.

References

[1] E. G. Agulla, E. A. Rúa, J. L. A. Castro, D. G. Jiménez, and L. A. Rifón, "Multimodal biometrics-based student attendance measurement in learning management systems," in *Proc. 11th IEEE Int. Symp. Multimedia*, 2009, pp. 699–704.

[2] H. Ç. Akakin and B. Sankur, "Robust classification of face and head gestures in video," *Image Vis. Comput.*, vol. 29, no. 7, pp. 470–483, 2011.

[3] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychol. Bull.*, vol. 111, no. 2, pp. 256–274, 1992.

[4] A. Arya, L. N. Jefferies, J. T. Enns, and S. DiPaola, "Facial actions as visual cues for personality," *Comput. Animation Virtual Worlds*, vol. 17, nos. 3/4, pp. 371–382, 2006.

[5] M. Bartlett, G. Littlewort, E. Vural, K. Lee, M. Cetin, A. Ercil, and J. Movellan, "Data mining spontaneous facial behavior with automatic expression coding," in *Proc. Verbal Nonverbal Features Human-Human Human-Machine Interaction*, 2008, pp. 1–20.

[6] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: Automatic assessment using short self-presentations," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 255–262.

[7] 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 Int. Conf. Weblogs Social Media*, 2011.

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

[9] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez, "Facetube: Predicting personality from facial expressions of emotion in online conversational video," in *Proc. 14th Int. Conf. Multimodal Interaction*, 2012, pp. 53–56.

[10] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez, "Hi Youtube!: Personality impressions and verbal content in social video," in *Proc. 15th ACM Int. Conf. Multimodal Interaction*, 2013, pp. 119–126.

[11] P. Borkenau and A. Liebler, "Trait inferences: Sources of validity at zero acquaintance," *J. Personality Soc. Psychol.*, vol. 62, no. 4, p. 645, 1992.

[12] S. W. Chew, P. Lucey, S. Lucey, J. M. Saragih, J. F. Cohn, I. Matthews, and S. Sridharan, "In the pursuit of effective affective computing: The relationship between features and registration," *IEEE Trans. Syst., Man, Cybern. B*, vol. 42, no. 4, pp. 1006–1016, Aug. 2012.

[13] J. Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce, and R. Schultz, "Smilemaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2008, pp. 978–986.

[14] S. Counts and K. B. Stecher, "Self-presentation of personality during online profile creation," in *Proc. Int. Conf. Weblogs Social Media*, 2009, p. 1.
[15] C. Darwin, *The Expression of the Emotions in Man and Animals*, P. Ekman, Ed. New York, NY, USA: Harper Perennial, 1872/2009.
[16] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Nat. Acad. Sci. USA*, vol. 111, no. 15, 2014, pp. E1454–E1462.
[17] G. Duchenne, and R. Cuthbertson. (1990). *The Mechanism of Human Facial Expression*, series Cambridge books online. Cambridge, U.K.: Cambridge Univ. Press [Online]. Available: http://books.google.es/books?id=a9tjQC7xbNMC
[18] P. Ekman, *Darwin and Facial Expression: A Century of Research in Review*. New York, NY, USA: Academic, 1973.
[19] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.
[20] P. Ekman and E. L. Rosenberg, Eds., *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*, 1st ed., series, Series in affective science. London, U.K.: Oxford Univ. Press, 1997.
[21] M. Frank, M. Bartlett, and J. Movellan, "The m3 database of spontaneous emotion expression (University Buffalo)," In pres., 2010.
[22] M. G. Frank, P. Ekman, and W. V. Friesen, "Behavioral markers and recognizability of the smile of enjoyment," *J. Personality Soc. Psychol.*, vol. 64, no. 1, p. 83, 1993.
[23] S. D. Gosling, S. Gaddis, and S. Vazire, "Personality impressions based on facebook profiles," presented at the *Int. Conf. Weblogs Social Media*, Boulder, CO, USA, 2007.
[24] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, "A very brief measure of the big-five personality domains," *J. Res. Personality*, vol. 37, pp. 504–528, 2003.
[25] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. on PAMI*, vol. 21, no. 10, pp. 974 –989, 1999.
[26] J. A. Hall, S. D. Gunnery, and S. A. Andrzejewski, "Nonverbal emotion displays, communication modality, and the judgment of personality," *J. Res. Personality*, vol. 45, no. 1, pp. 77–83, 2011.
[27] T. Kanade, Y. Tian, and J. F. Cohn, "Comprehensive database for facial expression analysis," in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2000, pp. 46–53.
[28] C. F. Keating, A. Mazur, M. H. Segall, P. G. Cysneiros, J. E. Kilbride, P. Leahy, W. T. Divale, S. Komin, B. Thurman, and R. Wirsing, "Culture and the perception of social dominance from facial expression." *J. Personality Soc. Psychol.*, vol. 40, no. 4, p. 615, 1981.
[29] M. L. Knapp and J. Hall, *Nonverbal Communication in Human Interaction*. New York, NY, USA: Holt, Rinehart and Winston, 2005.
[30] B. Knutson, "Facial expressions of emotion influence interpersonal trait inferences," *J. Nonverbal Behav.*, vol. 20, no. 3, pp. 165–182, 1996.
[31] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro, "Modeling the personality of participants during group interactions," presented at the 17th Int. Conf. User Model., Adaptation, Personalization,Trento, Italy, 2009.
[32] G. Littlewort, M. S. Bartlett, L. P. Salamanca, and J. Reilly, "Automated measurement of children's facial expressions during problem solving tasks," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, 2011, pp. 30–35.
[33] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (CERT)," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, 2011, pp. 298–305.
[34] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, and J. Cohn, "Aam derived face representations for robust facial action recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 155–160.
[35] S. Mariooryad, and C. Busso, "Feature and model level compensation of lexical content for facial emotion recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, 2013, pp. 1–6.
[36] D. Matsumoto and T. Kudoh, "American-Japanese cultural differences in attributions of personality based on smiles," *J. Nonverbal Behav.*, vol. 17, no. 4, pp. 231–243, 1993.
[37] R. R. McCrae and O. P. John, "An introduction to the five-factor model in its applications," *J. Personality*, vol. 62, pp. 175–215, 1992.
[38] D. McDuff, R. el Kaliouby, and R. Picard, "Crowdsourced data collection of facial responses," in *Proc. 13th Int. Conf. Multimodal Interfaces*, 2011, pp. 11–18.
[39] J. M. Montepare and H. Dobish, "The contribution of emotion perceptions and their overgeneralizations to trait impressions," *J. Nonverbal Behav.*, vol. 27, no. 4, pp. 237–254, 2003.
[40] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interfaces*, Nov. 2011, pp. 169–176.
[41] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, "Personality judgments based on physical appearance," *Personality Soc. Psychol. Bull.*, vol. 35, no. 12, pp. 1661–1671, 2009.
[42] B. O'Conaill, S. Whittaker, and S. Wilbur, "Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication," *Hum.-Comput. Inter.*, vol. 8, no. 4, pp. 389–428, Dec. 1993.
[43] A. Roy-Charland, M. Perron, O. Beaudry, and K. Eady, "Confusion of fear and surprise: A test of the perceptual-attentional limitation hypothesis with eye movement monitoring," *Cogn. Emot.*, vol. 28, no. 7, pp. 1214–1222, 2014.
[44] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. De la Torre, and A. Ross, "Automated facial expression recognition system," in *Proc. IEEE Int. Carnahan Conf. Secur. Technol.*, Oct. 2009.
[45] M. Sayette, J. Cohn, J. Wertz, M. Perrott, and D. Parrott, "A psychometric evaluation of the facial action coding system for assessing spontaneous expression," *J. Nonverbal Behav.*, vol. 25, pp. 167–186, 2001.
[46] K. L. Schmidt and J. F. Cohn, "Human facial expressions as adaptations: Evolutionary questions in facial expression research," *Amer. J. Phys. Anthropol.*, vol. 116, no. S33, pp. 3–24, 2001.
[47] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.*, vol. 86, no. 2, pp. 420–428, Jan. 1979.
[48] S. Trichas and B. Schyns, "The face of leadership: Perceiving leaders from facial expression," *Leadership Quart.*, vol. 23, no. 3, pp. 545–566, Mar. 2012.
[49] E. Vural, M. Cetin, A. Ercil, G. Littlewort, M. Bartlett, and J. Movellan, "Drowsy driver detection through facial movement analysis," in *Proc. IEEE Int. Conf. Human-Comput. Interaction*, 2007, pp. 6–18.
[50] J. S. Wiggins and A. L. Pincus, "Personality: Structure and assessment," *Annu. Rev. Psychol.*, vol. 43, no. 1, pp. 473–504, 1992.
[51] J. Willis and A. Todorov, "First impressions," *Psychol. Sci.*, vol. 17, no. 7, pp. 592–598, Jul. 2006.
[52] M. Wollmer, F. Weninger, T. Knaup, B. Schuller, K. Sagae, and L.-P. Morency, "Youtube movie reviews: In, cross, and open-domain sentiment analysis in an audiovisual context," *IEEE Intell. Syst.*, vol. 28, no. 3, pp. 46–53, Mar. 2013.

Lucía Teijeiro Mosquera received the master's degree in telecommunications engineering from the University of Vigo, Spain, in 2008, where she is currently working toward the PhD degree. She visited Idiap Research Institute and Bogazii University. Her research interests include computer vision, face processing, statistical modeling, facial expression recognition, and computer vision for driver assistance.

Joan-Isaac Biel received the PhD degree from the Swiss Federal Institute of Technology in Lausanne (EPFL) in June 2013. He carried out his doctoral research at the Idiap Research Institute, and has visited Yahoo! Labs, Barcelona, HP Labs, Palo Alto, and the International Computer Science Institute, Berkeley. His research is focused on the analysis of human communication, interaction, and multimedia engagement with online social video.

José Luis Alba Castro received the PhD degree in telecommunications engineering in 1997 from the University of Vigo, Spain, where he is currently an associate professor, teaching image processing, statistical pattern recognition, machine learning, and biometrics. His research interests include signal and image-based biometrics, computer vision for driver assistance, and computer vision for quality control. He is the head of the computer vision laboratory of the University of Vigo and has been the leader of many research projects and contracts on these topics. He has served in many technical program committees and authored more than 80 research papers. He is an associate editor of *EURASIP Journal on Information Security*.

Daniel Gatica-Perez is the head of the Social Computing Group at Idiap Research Institute and Maitre d'Enseignement et de Recherche at the Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland. His research interests include social computing, mobile and ubiquitous computing, and social media. He has served as an associate editor of the *IEEE Transactions on Multimedia*. He is a member of the IEEE.