The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs

Joan-Isaac Biel and Daniel Gatica-Perez, *Member, IEEE*

***Abstract*—Despite an increasing interest in understanding human perception in social media through the automatic analysis of users' personality, existing attempts have explored user profiles and text blog data only. We approach the study of personality impressions in social media from the novel perspective of crowdsourced impressions, social attention, and audiovisual behavioral analysis on slices of conversational vlogs extracted from YouTube. Conversational vlogs are a unique case study to understand users in social media, as vloggers implicitly or explicitly share information about themselves that words, either written or spoken cannot convey. In addition, research in vlogs may become a fertile ground for the study of video interactions, as conversational video expands to innovative applications. In this work, we first investigate the feasibility of crowdsourcing personality impressions from vlogging as a way to obtain judgements from a variate audience that consumes social media video. Then, we explore how these personality impressions mediate the online video watching experience and relate to measures of attention in YouTube. Finally, we investigate on the use of automatic nonverbal cues as a suitable lens through which impressions are made, and we address the task of automatic prediction of vloggers' personality impressions using nonverbal cues and machine learning techniques. Our study, conducted on a dataset of 442 YouTube vlogs and 2210 annotations collected in Amazon's Mechanical Turk, provides new findings regarding the suitability of collecting personality impressions from crowdsourcing, the types of personality impressions that emerge through vlogging, their association with social attention, and the level of utilization of nonverbal cues in this particular setting. In addition, it constitutes a first attempt to address the task of automatic vlogger personality impression prediction using nonverbal cues, with promising results.**

***Index Terms*—personality, video blogs, vlogs, YouTube, nonverbal behavior, multimodal analysis, crowdsourcing, interpersonal perception**

I. Introduction

Online video is more than YouTube, its 48 hours of video uploaded per minute [60], and all the singers, dancers, actors, journalists, and musicians contributing to the site. Conversational video is constantly evolving and expanding to new online applications with the purpose of making communication more natural, more effective, and more engaging: "My initial approach to this company was to create more intimacy in how people communicate, video was just the by-product." - says the founder of VYou, a question answering video platform startup [55]. This and other new platforms are driving online conversational video to innovative forms of interaction that address specific communication intents, such as chatting [1], dating [2], or sharing video testimonials for marketing [3], just to mention some.

J.-I. Biel and D. Gatica-Perez are affiliated jointly to the Idiap Research Institute, Martigny, Switzerland, and Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (e-mail: jibiel@idiap.ch; gatica@idiap.ch)

While all these emerging forms of video interaction are taking off, conversational video blogging (vlogging) has become a well established type of online conversational video, and one of the most prevalent formats among user-generated content in sites like YouTube [16]. In their typical, single talking-head setting, vlogs can be thought as a multimodal extension of text-based blogging, where words – what is said – are enriched by the complex nonverbal behavior displayed in front of the camera – how it is said. Though ethnographic studies have started to investigate these practices [37], conversational vlogging remains largely unexplored as a research subject. Conversational vlogs are a unique case study to understand users in social media, yet we do not know of many attempts to analyze them using automatic multimodal techniques that extract the behavior displayed and are applicable at large scale [11], [12].

Recent research in social networks and blogs has addressed the study of personality as a suitable lens for user understanding in social media. Based on user profiles, some research has suggested that the behavior of users in social networks reveals a clear intent of sharing their personality online [25], [20]. Some works have also shown that users can make accurate personality impressions from the information displayed in social network user profiles [25], and have investigated what specific features from user profiles and photos are more useful to create personality impressions [20], [22]. Regarding text blogs, some research has used text content analysis to investigate the association between personality traits and the use of specific words [24], [59], and has also addressed the automatic classification of personality from blog data with promising results [46], [45]. To our knowledge, however, no work has addressed the study of personality in vlogging.

We approach the study of personality impressions in vlogging from the perspective of nonverbal behavioral analysis. Nonverbal behavior is effectively used to express aspects of identity such as age, occupation, culture, and personality, and therefore it is also used by people to form impressions about others [5]. Moreover, nonverbal cues are useful to characterize social constructs related to human internal states, traits, and relationships as shown in social psychology [35] and social computing [47], [23]. This is relevant in the context of social media, not only because nonverbal behavior has been so far unexplored, but also because the nonverbal channel conveys information that is often unconscious and difficult to control [35], compared to the content people post in social networks and text blogs [25]. In this work, we investigate nonverbal behavior as a source of information that is used by (potentially large) audiences that watch videos online and create personality impressions about video bloggers

(vloggers). Following this social video-watching paradigm, we address four previously unexplored issues which are the main contributions of our work:

- We study the feasibility of collecting vlogger personality impressions using crowdsourcing techniques, thus going beyond the traditional laboratory based setting, in a framework that is suitable for the annotation of large-scale data, and that resembles the ways in which online video is consumed in sites such as YouTube, where ordinary, diverse people (as opposed to trained annotators) make first impressions while they watch thin-slices of video.
- We study how personality impressions mediate the vlog watching experience, investigating the associations between crowdsourced personality impressions and the actual levels of attention that vloggers gather online, using several measures of attention estimates from YouTube's available metadata.
- We approach the problem of individual cue utilization to investigate to what extent nonverbal cues extracted automatically from audio and video are useful as a lens through which personality impressions from vlogs can be made.
- We address the task of automatic predicting vloggers' personality impressions using multimodal nonverbal cues and machine learning techniques.

Some preliminary results of this work appeared in [10]. In the current manuscript we provide a comprehensive, full presentation of the methodological and analytical context in which we ground our work, providing explanations of our findings regarding the suitability of collecting personality impressions fro crowdsourcing, the types of personality impressions that emerge through vlogging compared to other human interaction settings, and the association of vloggers' behavior with social attention measures from YouTube. In addition, it constitutes a first attempt to address the task of automatic vloggers personality impression prediction using nonverbal cues, with promising results.

The rest of this paper is structured as follows: Section II introduces related work. Section III outlines our approach on the study of personality impressions and nonverbal behavior in vlogging. Section IV presents our data collection and preprocessing techniques. Section V describes our personality impressions collection using Amazon's Mechanical Turk. Section VI describes the feature extraction process. Section VII presents and discusses the results on each of the four points introduced. Finally, we conclude in Section VIII.

II. Related work

We first review some basic concepts on personality research that define our framework. Then, we review related work on studying personality in social media; collecting human judgments for annotation of multimedia corpora; the study of social attention in social media, and recognizing personality automatically from audio and video.

A. Personality Research and the Big-Five Model

The Big-Five framework of personality is a hierarchical model that organizes personality traits in terms of five basic bipolar dimensions: Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to Experience (O). These five dimensions are easily interpreted by referring to their associated personality attributes, as presented in Table I [42]. Though the Big-Five model has not been universally accepted, it has considerable support and has become the most widely used and researched model of personality [26]. To measure the extent to which each one of these traits describes human personality, several rating instruments have been developed, which are used to ask people to rate themselves (a.k.a. self-reported personality) or to rate others based on the raters' self-observations (a.k.a. personality impressions). Self-reported personality and personality impressions provide two different views of personality perception and have been used to answer questions such as:

1) Do observers agree on their personality impressions?
2) Are personality impressions accurate compared to self-reported personality?
3) How much information is needed to achieve large agreement (or accuracy) in personality impressions?

Regarding 1) and 2), research has shown that external observers agree on their personality impressions of targets, and that such impressions are fairly accurate with the targets' self-reported personality, even if impressions are formed in the presence of minimal information, in several contexts: physical presence [4], video-tapes [13], [33], photos [13], websites [56], bedrooms [27], etc. Regarding 3), research has investigated the extent to which certain appearance or behavioral cues convey information associated to personality impressions (cue utilization) and to the actual self-reported personality (cue validity), and has shown that the nature of valid and useful cues depends on the context in which impressions are made [27], [13].

In this paper, we are interested in the investigation of vloggers personality impressions and the utilization of automatic nonverbal cues extracted form audio and video. However, given the relatively small amount of works addressing the study of personality in both social media and social computing, we also review some works addressing the use of self-reported personality traits, whenever we found them relevant from a computational perspective.

B. Personality Research in Social Media

Our work relates to social media research that addresses the study of personality impressions form social network users' profiles [25], [22], [53], [20], [52]; the use of automatic content analysis to study bloggers personality [24], [59], [45], and the prediction of user personality from blog data [46], [45].

In one of the first studies of personality in social media, Gosling et al. [25] showed that personality impressions inferred from user profiles achieved significant agreement among observers, and that these impressions are accurate compared to the self-reported personality of the profile owners. They also found evidence that user profiles may reflect the actual personality of their owners, rather than a self-idealized one [25]. Also

TABLE I
BIG FIVE PERSONALITY TRAITS AND ASSOCIATED ADJECTIVES [42].

Personaliy Trait	*Adjectives*
Extraversion (E)	Active, Assertive, Energetic, Enthusiastic, Outgoing, Talkative
Agreeableness (A)	Appreciative, Forgiving, Generous, Kind, Sympathetic
Conscientiousness (C)	Efficient, Organized, Planful, Reliable, Responsible, Thorough
Neuroticism* (N)	Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying
Openness to Experience (O)	Artistic, Curious, Imaginative, Insightful, Original, Wide Interests

*Neuroticism may alternatively be presented as Emotional Stability by inverting the scale.

in the context of user profiles, Evans et al. [22] and Steele et al [53] identified what elements from user profiles and photos are associated to personality impressions being more or less accurate. Furthermore, Stecher and Counts [52], [20] studied the process of profile creation in relation with both the memory of personality impressions [52] and the convergence between self-assessed personality and desired self-presentation [20]. Compared to our research, none of these works involved any automatic analysis of multimedia content.

Research in social media has also used automatic content analysis of text blogs to extract cues from word usage and writing styles [24], [59], [45]. Gill et al. [24] examined a sample of 2,400 blogs and 12M words to study the relation between people self-reported personality and word usage (cue validity), and the relation between their personality and their motivations for blogging. In a larger sample of blogs (700 users and 80M words), Yarkoni [59] also addressed the problem of cue validity, providing a more exhaustive study of the links between both word usage and linguistic styles and bloggers' personality.

Some initial research has also addressed the task of automatic prediction of personality in the context of text blogs [46], [45]. In a first attempt, Oberlander and Nowson [46] investigated the use of an n-gram model to automatically classify the self-reported personality of 70 bloggers, achieving accuracies between 75% and 84% on a balanced, binary classification task. However, when the same approach was used to classify 1769 blogs accuracies, decreased down to 52% and 59% [45]. The authors highlighted the difficulty of classifying personality using noisy, large-scale data from blogs and personality scores with relatively low reliability (whereas the questionnaire used in the first experiment included 41-items, the second questionnaire included only 5 items). Independently of whether the authors were interested on predicting the self-reported personality of bloggers rather than the impressions that readers made about them, these works show the potential of using automatic techniques for the prediction of personality in large-scale social media data.

Compared to these works, we focus in a less studied social media community, namely vloggers, with the goal of providing new insights about the creation and perception of personal online videos. In addition, we investigate the use of automatic nonverbal behavioral cues as an information source present in social media that is useful to study users. This is relevant, not only because nonverbal behavior has been unexplored up to date in the context of social media, but also because the nonverbal channel conveys information that is often unconscious and difficult to control [35] compared to the content people post in social networks and text blogs [25].

C. Crowdsourcing Human Impressions for Multimodal Corpora Annotation

Along the line of recent literature showing the utility of micro-tasks markets for conducting human behavioral studies [34], our work contributes to recent efforts that explore the feasibility of crowdsourcing the annotation of multimodal corpora based on human observations [51], [14].

The study of personality impressions requires the collection of personality observations from a third-party. In some works, these observations can be collected between participants during the studies, specially on face-to-face interactions [4]. However for the study of personality impressions using multimodal corpora, a common approach consists on gathering people and ask them to listen/watch the data, and to judge the personality of the people on them [40], [43]. This approach, which is used in general to annotate other personal and social constructs in data corpora [32], can become expensive in terms of time and money, specially if one aims to annotate large-scale data.

Recently, Soleymani and Larson [51] explored the use of Amazon's Mechanical Turk crowdsourcing platform to annotate the affective response of people to a set of 126 videos. They concluded that crowdsourcing was a valuable technique to collect affective annotations and provided a short guide of best practices to use Mechanical Turk for that purpose. In another work, Brew et al. [14] crowdsourced the annotation required to train a machine learning system to score and track news feeds' sentiments, and investigated how to manage the effort of annotators to maximize the coverage and the agreement achieved in the annotations. Clearly, these tasks differ from other crowdsourced annotations in that there is no clear cut ground truth data. Related to the collection of personality data from crowdsourcing, Buhrmester et al. [15] explored the collection of data for psychology studies by administering a series of self-reported personality questionnaires to MTurk workers, and found that the quality of data met the psychometric standards associated with published research. In addition, the results suggested that MTurk participants are at least as diverse and more representative of non-college populations than those of typical Internet and traditional social psychology samples. This is indeed an interesting feature of MTurk for our collection of personality impressions, because our purpose is to collect observations made by ordinary people (as opposed to trained annotators).

To our knowledge, our work constitutes a first attempt to crowdsource personality impressions from online video in a framework that is suitable for the annotation of large-scale data, and that resembles a diverse online community who watches appealing content and disregards uninteresting one.

D. Personality Impressions and Social Attention

Our work also contributes to existing literature that has emphasized the importance of human attention as a social,

valuable good sought for everyone in multiple social contexts [29], and that has discussed several personal, social, and behavioral aspects related to achieving attention or reacting to it, both online [28], [12] and offline [7], [6], [13].

Recent research has investigated the role of attention in the production and consumption of content within new digital media [28], [29]. Regarding YouTube, for example, Huberman et al. [29] showed that the productivity of users uploading videos strongly depends on attention (as measured by the number of views videos received) and that, in addition, lack of attention leads to a decrease in the number of videos uploaded and a consequent drop in productivity. In a more recent study on YouTube conversational vloggers' nonverbal behavior [12], we found evidence that some audio, visual, and multimodal nonverbal cues extracted from vlogs are correlated with this same measure of attention. Without claiming any causality effects between vloggers' nonverbal behavior and attention, this result suggests that vloggers receiving similar attention share personal characteristics that are manifested in their nonverbal behavior, and that underline how effective they are at communicating and creating vlogs.

In this work, we address this problem from the lens of vloggers' personality as high-level, broad descriptors of other characteristics related to human communication such as persuasion, social skills, status, respect, creativity, etc. More specifically, we hypothesize that some personality impressions from vloggers are correlated with the level of attention that the vloggers achieve, somehow revealing the personality traits that are most appealing to people during video watching. In this line, our work is related to a number of research works that investigate the ways in which certain personality traits attract social attention [7] , and to works investigating the correlates of personal characteristics and appearance with social outcomes such as respect, status [6], or popularity in social groups [13].

E. Automatic Modeling of Personality from Audio and Video

Finally, our work adds to recent literature that models personality computationally using automatic nonverbal behavioral cues from audio and video in several face-to-face communication scenarios, and that is inspired on much classic social psychology works on personality and nonverbal behavior [48], [21], [31], [5], [35].

Mairesse et al. [40] investigated the prediction of personality impressions from 96 individual segments of conversational audio. Their regression using audio cues only resulted in performances of R^2 = 24%, 18%, and 15% for Extraversion, Emotional Stability, and Conscientiousness respectively. This work emphasized the role of prosodic cues in the automatic prediction of personality. In a related study, Mohammadi et al. [43] focused on the use of prosodic cues to automatically classify Big-Five personality impressions obtained on French professional radio broadcasts (7h of data). The obtained accuracy ranged from 65% to 80% depending on the traits, and the authors reported the Extraversion trait as the easier to predict.

In the context of small group meetings, Lepri et al. [38], [39] investigated the automatic prediction of self-reported extraversion and locus of control. They explored the use of audio cues (speech activity and prosody) and visual cues (energy from head, hands, and body) extracted on one-minute slices in a total of 6h of data. In their regression task, they obtained up to R^2 = 22% for Extraversion [38]. They also found that using the gaze of others as a cue could help to improve the prediction [39]. Typically related to Extraversion, dominance has also computationally modeled using automatic nonverbal cues [30], [32]. Jayagopi et al. [32] investigated audio-visual cues for dominance estimation using both an unsupervised and supervised model and found that speaking time led to superior performance. Also in group meetings, a high ratio between looking while speaking and looking while listening [21] was found to determine dominance [30].

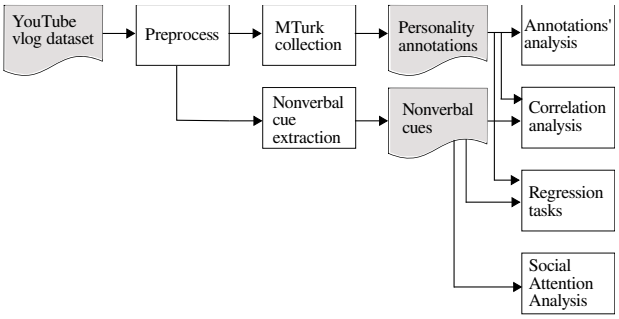

Fig. 1. Overview of our approach for the study of personality impressions in YouTube vlogs using nonverbal cues.

Though vlogs are not face-to-face interactions, vloggers behave in ways that that resemble having a conversation with their audience through their webcam [58]. Thus, in this work we investigate the suitability of using similar conversational nonverbal cues in the context of vlogging. At the time of publication, we found a piece of work that addressed the study of personality from automatic audiovisual analysis in a similar monologue setting than vlogging [9]. We found that whereas the techniques used for analyzing such content may be similar than the ones used here, the data used and the task addressed are completely different in nature. First, as opposed to the recordings made in a laboratory [9], our data consists of spontaneous self-recorded vlogs that are made to be shared publicly online. Second, we address the problem of personality impressions and not self-reported personality.

III. Overview of Our approach

Figure 1 summarizes the technical blocks of our approach for the study of personality impressions. We start by preprocessing a set of conversational vlogs from YouTube to create "thin-slices" of behavior with the extraction of the first conversational minute. On one hand, we use these vlog slices to obtain vloggers' personality judgments from Mechanical Turk workers. On the other hand, we process the slices to automatically extract nonverbal cues from audio and video. Then, we divide the study in four different blocks. First, we analyze the work of MTurk annotators and study the agreement of personality impressions. Second, we investigate the relation that these impressions have with social attention, measured from YouTube metadata. Third, we measure the level of cue utilization of automatic nonverbal cues from vloggers as lenses

that mediate the personality impressions of observers. Finally, we address the task of predicting personality from vloggers' from automatically extracted nonverbal behavior. Each module is discussed in Sections IV-VIII.

IV. YouTube Vlog Dataset and Preprocessing

Our set of YouTube vlogs consists of 2,269 videos between 1 and 6 min long from 469 different vloggers (a total of 150h of video), together with all available video metadata and viewers' comments were collected in 2009 [11]. The sample had been obtained using keywords such as "vlogs" and "vlogging" and had been limited to a setting in which a single person was talking to the camera mostly showing head and shoulders. No other restriction was imposed in terms of content or topics, which are variate, including personal vlogs, movie, books, and product reviews, political debate, etc.

With the purpose of bounding the time needed to collect human annotations for our personality impressions collection, we limited the size of the dataset for our experiments to one video per vlogger, and we preprocessed the videos to obtain the first conversational minute only. The use of "thin slices" is greatly documented in psychology as suitable for the study of first impressions [5], [47]. For the case of personality, some research has suggested that few seconds are enough to make accurate judgments [17].

Vlogs may result from a combination of conversational and non-conversational shots mixed in any order. For example, non-conversational shots at the beginning of the video may be used as user-branded opening to a vlog [12]. Recently, we showed how to automatically obtain useful conversational/non-conversational shot segmentations with the use of a shot boundary detector and an automatic face detection system [12]. In this work, we used the same approach to process every vlog. Then, we either shortened the first conversational shot or merged it with subsequent conversational shots when required, in order to obtain the first conversational minute. In practice, we allowed durations between 50s and 70s so to minimize the number of shot boundaries in the final vlog slice, and we discarded 27 vlog slices that resulted shorter than 50s after merging the conversational segments. The final dataset contained 442 vlogs of which 47% (208) corresponded to male and 53% (234) to female vloggers.

V. Using Mechanical Turk to Collect Personality Impressions

We addressed the task of collecting personality impressions using Amazon's Mechanical Turk, following a video-watching paradigm that aims to resemble the ways in which online video is consumed in sites such as YouTube, where ordinary, diverse people (as opposed to train users) make first impressions while they watch thin-slices of video. Using MTurk, we also intended to explore the possibility of an affordable and fast completion method that could truly scale to the annotation of large amounts of social media data.

Figure 2 shows a snapshot of our Human Intelligence Task (HIT) design, which consisted of two main components. The top part of the HIT contained an embedded video player to

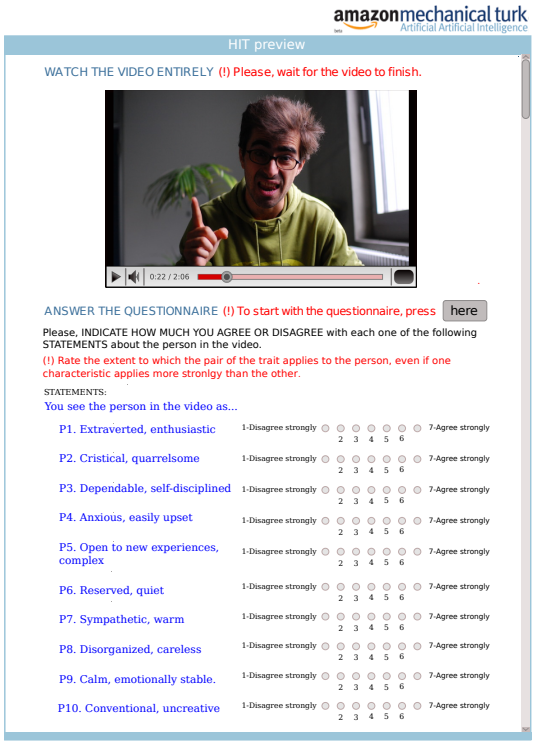

Fig. 2. A view of the HIT designed to collect personality judgments from MTurk. On the top, the embedded vlog. On the bottom, the TIPI questionnaire.

display the one-minute vlog slices obtained from preprocessing (see Section IV). The bottom part of the HIT included the Ten-Item Personality Inventory form designed by Gosling et al. [26]. This personality inventory was designed to measure personality in settings like ours, were there are limitations on the time that people can spend to complete the questionnaires, and measures the Big Five personality traits on the basis of only 10 items (two items per scale on a 7-point likert scale). This is also the case in our setting, because we aim to keep the task short. With the purpose of obtaining spontaneous impressions, we did not give any particular instructions to workers to fill the questionnaire apart form 1) watching the video entirely and 2) answering the questionnaire. The TIPI instructions were directly taken from [26], but were rephrased to ask about the vlogger personality. We also added an extra third component to the HIT to collect some demographics from the vloggers: the gender, the age, and the ethnicity. Age and ethnicity were divided in six categories each: younger than 12, 12-17, 18-24, 25-34, 35-50, and older than 50 (for age), and Caucasian, Black or African American, Asian/Pacific Islander, American Indian/Alaskan native, Hispanic, and Other (for ethnicity). These questions represent an opportunity to measure the reliability and quality of the MTurk annotations, as they are typically clearly more objective than the personality questionnaire.

The actually design of the HIT, resulted from an iterative process, in which we conscientiously refined the HIT design to discourage spammers from completing our tasks. With this purpose, we incorporated javascript and CSS to disable the HTML questions and control the flow of the HIT. First, the questionnaire was enabled only after the video had reached

the end. Second, the demography questions were enabled only after all the TIPI was completed to make sure that no items were skipped. Third, the final "Submit" button of the MTurk interface was hidden until all the questions were answered. Finally, in addition to the working time reported by MTurk, we logged the time spend on each of the components: time watching the video, time filling the TIPI questionnaire, and time answering the demographic questions.

Gosling et al. [26] suggested that the TIPI could be completed in one minute. Thus, we estimated each HIT to take no more than 2 minutes. We posted a total of 2,210 to collect five different judgments for each of the 442 vloggers. The HITs were restricted to workers with HIT acceptance rates of 95% or higher, from the US (1,768 HITs) and India (442 HITs), as these are the English speaking countries with more MTurk workers [41].

VI. Nonverbal Cue Extraction

We automatically processed the one-minute vlogs to extract nonverbal cues from both audio and video as descriptors of the vloggers' behavior. Given the conversational nature of vlogs, most of the features computed here were borrowed from social psychology and social computing works on conversational interactions [47], [23]. In addition, as suggested by their performance on the automatic prediction of personality [40], [43], we considered a broad range of prosodic nonverbal cues. Overall, these features result from aggregates along the video, i.e., for each vlog each cue is represented with a single value. The following sections provide a basic description of these cues and their computation.

A. Audio cues

We characterized the audio modality of vlogs with a set of nonverbal cues that measure patterns of speech activity, as well as prosody.

1) Speaking activity: We extracted speaking activity cues using the toolbox developed by the Human Dynamics group at MIT Media Lab [47]. These cues are extracted on the basis of a two-level hidden Markov model (HMM) that is used to segment the audio in voiced/unvoiced and speech/non-speech regions (Figure 3, top). From there, several cues measure how talkative and fluent people are.

- The speaking time (*Speaking Time*) is a measure of how much the vlogger talks. Though our vlogs display a monologue setting, we hypothesize that some vloggers may be more talkative than others depending on their personality. This feature is computed by the ratio between the total duration of speech and the total video duration.
- The length of the speaking segments (*Avg Length of Speak Segs*) is a measure of fluency, typically related to the duration and number of silent pauses (long segments are associated with short and few pauses) [48]. It is measured by the ratio between the overall duration of speech divided by the number of speech segments.
- The number of speaking turns (*# Speech turns*) is another measure of fluency, directly related to the number of silent pauses (silent pauses interrupt and initiate speaking turns) [47]. It is obtained by the ratio between the number of speech segments and the duration of the video.

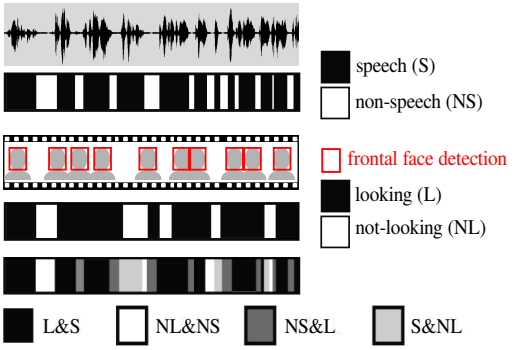

Fig. 3. Scheme of multimodal segmentations obtained from vlogs. Top: the speech/non speech segmentation obtained from audio using a two-layer HMM. Center: looking/not-looking segmentations obtained from video using a face detection system. Bottom: multimodal segmentation obtained by combining the audio and visual segmentations. (L = Looking, S= Speaking, N = Not).

2) Prosodic cues: Several prosodic cues are also obtained from audio. Voicing rate and some related cues were obtained form the MIT audio toolbox in a frame-by-frame basis (with windows of 32ms and steps of 16ms). In addition, energy and pitch features were obtained from the audio signal using PRAAT in windows of 40ms and time steps of 10ms. Finally, all the cues were aggregated between frames by computing the mean, median, mean-scaled standard deviation, maximum, minimum, and entropy.

- The voicing rate (*Voice rate*) relates to the frequency of phonemes while speaking, and represents the pace of a conversation [48]. In monologues, it can represent a measure of fluency or excitement. It is computed by the ratio between the overall duration of voice segments and the duration of speech. In addition to the voicing rate, we use the number of autocorrelation peaks (*# R0 peaks*), their location (*Loc R0 peaks*) and the spectral entropy (*Spec Entropy*) which are the raw features used to obtain the voicing/non-voicing segmentation [8].
- The speaking energy (*Energy*) is a measure of loudness, typically related to excitement. In its mean-scaled standard deviation form, it is also used to measure vocal control (how well the vlogger controls loudness), a feature typically related to emotionality [47]. We also computed the first derivative of the Energy (*D Energy*).
- The pitch (*F0*) is the main frequency of the audio signal. In its mean-scaled standard deviation form, it is another measure of vocal control and emotionality [47]. In addition to the pitch, we obtained the pitch bandwidth (*F0 BW*), intensity, (*F0 Intensity*), and the confidence of the estimate (*F0 Conf*), which were respectively aggregated as mentioned above.

B. Video cues

We characterized the video modality of vlogs by extracting nonverbal cues that capture patterns of looking activity, pose, and visual activity.

1) Looking activity and pose: Looking activity cues were extracted on the basis of a frontal face detection system [57] that was used to create binary segmentations of intervals looking/non-looking to the camera, under the sensible assumption that frontal face detections occur when the vlogger looks at the camera (see Figure 3). Though robust tracking methods could be applied here, this method does not require manual initialization nor parameter tuning, and it is robust regarding lighting variations and image resolution, and thus suitable for large-scale analysis. In addition, the face detection bounding box can be used as a proxy to measure proximity to the camera and framing.

- The looking time (*Looking Time*) is a measure of how much the vlogger looks to the camera. We hypothesize that despite the clear communication intent of vlogs, vloggers with different personalities may differ on the overall time spent looking to the camera. It is measured by the ratio between the overall looking time and the duration of the video.
- The length of the looking segments (*Av Length Look Seg*) is a measure of the persistence of a vlogger's gaze. It is computed by the ratio between the overall looking duration and the number of looking segments.
- The number of looking turns (*# Look turns*) measures how frequently the vlogger looks to the camera and it is directly related to patterns of gaze avoidance. It is obtained by the ratio between the number of looking segments and the video duration.
- The proximity to the camera (*Proximity to camera*) characterizes the choice of the vlogger to address the camera from a close distance. It is computed as the inverse of the average face bounding box area normalized by the video frame area.
- The vertical framing (*Vertical Framing*) measures to what extent faces are positioned on the upper part of the video frame and it is associated with vloggers showing the upper body. It is measured as the average vertical distance between the center of the bounding box and the center of the video frame normalized by the video frame height.

2) Visual activity: The overall motion of the vlogger is an indicator of vloggers' excitement and kinetic expressiveness. We computed the overall visual activity of the vlogger with a modified version of motion energy images called "Weighted Motion Energy Images" (wMEI) [10]. The normalized wMEI describes the motion throughout a video as a gray-scale image, where the intensity of each pixel indicates the visual activity on it. From the normalized wMEIs, we extract statistical features as descriptors of the vlogger's body activity such as the mean, median, and center of gravity (in horizontal and vertical dimensions). To compensate for different video sizes, all images are previously resized to the same dimensions (320x240 pixels).

C. Multimodal cues

We combined the speech/non-speech and looking/not-looking segmentations in one single audiovisual segmentation that captures regions of looking-while-speaking (L&S), looking-while-not-speaking (L&NS), and their counterparts (see Figure 3). We then adapted the dominance ratio designed for face-to-face interactions [21] to the vlog setting by computing the multimodal ratio L&S/L&NS. Our set of multimodal cues includes the multimodal ratio, as well as the individual values of L&S and L&NS.

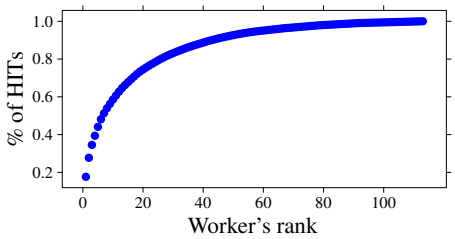

Fig. 4. Cumulative percentual distribution of MTurk annotations. Workers are ranked based on the number of HITs completed. The top ranked worker completed 17% of the HITs (400 HITs), 26 workers contributed with 80% of the annotations (1790 HITs), whereas 57 workers contributed with less than 5 HITs each (126 HITs).

VII. Results and Discussion

We divide our study in four parts. First, we analyze the work of MTurk annotators and measure the agreement of personality impressions. Second, we analyze the personality impressions in the context of YouTube, and investigate their correlations with social attention, as measured from vlogs' metadata. Third, we measure the utilization of automatic nonverbal cues from vloggers as lenses that mediate the personality impressions of observers by means of pair-wise correlations. Finally, we address the task of predicting personality from vloggers' nonverbal behavior using regression.

A. Crowdsourced Personality Judgements' Quality

The annotation tasks were completed by a total of 118 workers. The 1,760 HITs restricted to the US were completed by 91 workers and were finished within 12 days after being uploaded to MTurk, whereas the 442 HITs restricted to India were completed by only 27 workers and took 2 more days to be finished. This timing seems reasonable given the 75h hours of expected work (2min per HIT), the money spend on it, and the effort that would have taken to gather people offline to perform the task. However, it is slower than timings reported for other MTurk tasks [41]. Regarding individual work, each worker completed an average of 20 HITs. A two-tailed paired t-test on the distributions of HITs/worker showed no significant differences between the contribution of US and Indian workers ($t = -0.87$, $df = 103.021$, $p = 0.38$). However, as shown in Figure 4, the contribution varied substantially among workers, with one worker contributing to 17% of the annotations, and 26% of the workers providing up to 80% of the annotations. The average time of the TIPI questionnaire completion alone was 36.1s. Though this figure is substantially lower compared to the one minute suggested by Gosling et al. [26], this result agrees with other recent studies in MTurk, where completion times of annotations were reduced with respect to experts' working time, which can be justified by the economic motive of MTurk workers [51].

TABLE II
DESCRIPTIVE STATISTICS, PAIR-WISE CORRELATIONS, CRONBACH'S ALPHA, AND INTRACLASS CORRELATION COEFFICIENTS FOR PERSONALITY IMPRESSIONS ($^{***}p < .0001$).

	Trait	$Mean$	SD	$Skew$	Min	Max	1	2	3	4	5	α	$ICC(1,1)$	$ICC(1,K)$
1	Extr	4.61	1.00	-0.32	1.90	6.60		.03	.00	.08	$.56^{***}$	.58	$.39^{***}$	$.76^{***}$
2	Agr	4.68	0.87	-0.72	2.00	6.50			$.39^{***}$	$.69^{***}$	$.28^{***}$	.46	$.27^{***}$	$.64^{***}$
3	Cons	4.48	0.78	-0.32	1.90	6.20				$.55^{***}$	$.26^{***}$	.63	$.14^{***}$	$.45^{***}$
4	Emot	4.76	0.79	-0.57	2.20	6.50					$.31^{***}$	.61	$.13^{***}$	$.42^{***}$
5	Open	4.66	0.71	-0.09	2.40	6.30						.48	$.15^{***}$	$.47^{***}$

To get an idea of the level of impression agreement among workers, we first investigated the reliability on the three questions regarding the demographics of the vloggers by means of the Fleiss' Kappa coefficient. The Fleiss Kappa assesses the reliability of categorical ratings and compensates for the agreement that could occur if raters were annotating at random. As one would expect from a well performed task, the gender annotations showed high agreement ($\kappa = 91$). Clearly, the age and ethnicity annotations are more difficult to perform, yet they achieved a fair agreement for age ($\kappa = 29$), and moderated agreement for ethnicity ($\kappa = 46$).

Table II gathers some basic statistics regarding the actual personality impressions collected with the TIPI questionnaires. The personality impressions score of each vlogger were aggregated across the corresponding 5 available annotations computing the mean. All the personality traits resulted centered on the positive part of the likert scales (≥ 4) and showed little skewness (≤ 1). The table also shows the correlations between the impressions for the Big-Five traits (see columns labeled with 1-5), which was larger between Agreeableness and Emotional Stability ($r = .69$), and between Extraversion and Openness to Experience ($r = .51$).

We evaluated the quality of the MTurk annotations in terms of the internal consistency of the personality test. The TIPI questionnaire is expected to have lower consistency compared to other personality inventories, as a consequence of using only two items per scale [26]. However, very low consistency could indicate that MTurk workers are unreliably using the scales to fill up the personality questionnaires. To investigate this, we computed Cronbach's alpha reliability coefficient for each personality trait across all the annotations (N = 2,210), and report values in Table II. We obtained alphas between .46 and .63 depending on the traits, a value range similar to that reported on the original TIPI report [26], suggesting that MTurk workers are answering the questionnaires consistently from the point of view of the experimental design.

A cuore question of interest is to what extent workers are able to achieve any agreement on the basis of watching 1min slices of vlogs. In our setting, no agreement could result from two hypothetical situations in which either a) vloggers' behavior would not convey any personal information or b) MTurk workers did not pay attention while completing the HITs. We computed the Intraclass Correlation Coefficients (ICCs) for each personality trait, as they are commonly used in psychology to measure the level of absolute agreement between annotators [49]. Note that contrary to other existing reports of annotators agreement of personality [25], [56], we cannot use ICC(2,1) and ICC(2,k) measures, because each observer only annotated a subset of the data. Instead, we computed ICC(1,1) and ICC(1,k) which are measures of absolute agreement designed for experimental settings were each target is annotated by a k judges randomly selected from a population of K judges, with $k < K$ [49]. In our setting, we have $k = 5$ and $K = 113$.

We present the two ICCs for each trait in the last two columns of Table II. The ICC(1,1) measures the extent to which two perceivers agree with each other. In addition, the ICC(1,k) measures the degree of agreement in rating the targets, when the annotations are aggregated across the 5 workers to obtain a unique aggregate personality score. The ICC(1,1) shows moderate to low reliabilities for the single MTurk annotations (.15 $<$ ICC(1,1) $<$.40, $p < 10^{-3}$), whereas the ICC(1,k) display moderate reliabilities for the aggregated annotations (.47 $<$ ICC(1,k) $<$ .77). We remark a few observations in the context of previous personality impressions research. The first one is that different personality traits achieved substantially different agreement. The second is that Extraversion is the trait achieving the highest level of agreement among observers. These two results have been repeatedly reported in research in personality [25], [4], and are typically related to the evidence that the amount of observable information associated with some personality traits is larger than for others, and that this information varies with the context in which personality impressions are formed [27], [13]. Third, most of the literature in face-to-face and videotaped impressions consistently reported Conscientiousness as the trait showing the second highest reliability. Thus, it is very interesting to see that in our case the trait achieving the second highest ICC is Agreeableness, and not Conscientiousness (which is in fact the trait with second lowest ICC). As argued by Gosling et al. [27], who found a similar effect with the Openness to Experience trait in personality impressions from bedrooms, this may well indicate that the vlogging setting is providing much more valuable information to form impressions of Agreeableness, compared to information regarding Conscientiousness.

We shall emphasize that the magnitude of these reliabilities compares well to other personality impression works, and indicate that there is substantial agreement on the personality impressions from MTurk. For example, Ambady et al. [4] found that single personality impressions based on face-to-face interactions achieved a reliability between .07 and .27 for different traits, whereas Gosling et al. measured reliabilities between .23 and .51 for single impressions from bedrooms using the same TIPI questionnaire [27]. Because these reliabilities were reported in terms of mean pair-wise correlations between raters, we computed these measures on our data for comparison (we considered only those annotators with

TABLE III
DESCRIPTIVE STATISTICS AND PAIR-WISE CORRELATIONS FOR YOUTUBE ATTENTION MEASURES (ALL CORRELATIONS HAVE $p < .0001$).

	Measure	*Mean*	*SD*	*Skew*	*Min*	*Max*	1	2	3	4	5
1	# Views	288.91	8.18	0.98	1	2406284		.86	.88	.85	−.28
2	#Times favorited	2.07	3.66	1.93	0	22007			.90	.86	−.18.
3	# Raters	14.71	4.71	1.27	1	27349				.92	−.20
4	# Comments	12.50	5.31	0.99	0	23112					−.17
5	Average rating	4.85	0.87	−0.76	1	5					

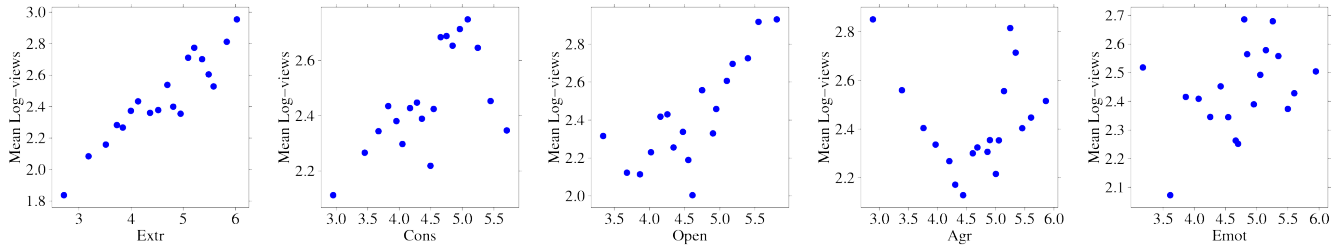

Fig. 5. XYplots for personality impressions and social attention based on # views received. All relations are linear except for Agreeableness, which is U-shaped.

more than 5 completed HITs). The resulting mean pair-wise correlations are: Extraversion (.44), Agreeableness (.36), Conscientiousness (.23) , Emotional Stability (.27), and Openness to Experience (.24). Other interesting works studying user profiles and websites reported reliabilities in terms of ICC(2,1) and may not be compared directly [56], [25].

Overall, our experience with MTurk was that running a HIT required substantially more involvement than we had estimated, answering emails from workers, ensuring that workers understand the task and take enough time, and validating submitted HITs in order to built a community of trusted workers. In general terms, our experience using MTurk supports that reported by others on the annotation of multimodal corpora [51]. However, this effort seems to be recompensed with annotations that have substantial agreement.

B. Personality Impressions and Social Attention

Several forms of social participation take place around vlogs. People may watch a vlog, but may also actively "like" or "unlike" it – manifesting some kind of approval or disapproval –, mark it as one of their favorites, or write a comment about it. All these actions, which are registered and aggregated by the YouTube platform in the form of metadata, can be interpreted as signaling the attention that vloggers receive from the YouTube audience. Because each one of this actions involves people to a different degree, one could argue that these different actions unveil different aspects of people's attention. For example, anybody can watch videos in YouTube, but only registered and logged-in people can "like", "favorite", or comment in a vlog. Moreover, writing a comment clearly takes more time and effort than just "liking" the video. In practice, however, most of these measures are largely correlated [12].

Table III summarizes some basic statistics of these measures as obtained from YouTube metadata for the 442 vlogs in our dataset. The # views, # time favorited, # raters and # comments distributions are highly skewed to very large values, due to the long-tail distribution of social network data, where a small percentage of vloggers are very popular and lots of them are ordinary. To reduce skewness, we transformed these measures with a log function. Note that the mean, standard deviation, minimum, and maximum values shown in the table were computed in the log scale and were transformed back for displaying purposes. Only the skewness measure was computed after transformation. The case of the average rating is quite different. This measure, which ranges between 1 and 5, showed a large bias towards large values (mean = 4.85), which suggests that when people decide to rate a vlog tend to give high ratings. We transformed this measure with a power ten function. We also computed the inter-class correlations between these measures, which are shown in the right part of Table III. Note that only the correlations on the last column are relatively low, which indicates that the average rating received by the videos is weakly related to the # of views, # times favorites, # raters, and # comments of videos.

We investigate the personality correlates with aggregates of these measures of attention computed across groups of vlogs featuring similar personality scores. This aggregation method is used to compensate the attention measures from network processes such as preferential attachment [18], which inflate these measures, but may not result from a manifest interest of audiences in the content of the videos itself. Following a recent work by Huberman et al. [29] we compute the average level of attention as an aggregate of the attention received by a set of videos. Formally, this quantity is defined as $\hat{a} = \sum_{n=1}^{N} a_n/N$, where $\hat{a}$ is the average level of attention, N is the number of videos in the set, and a_n is the attention received by the n-th video as measured by one of the attention measures introduced above. The analysis proceeds as follows. For each personality trait, we divide vloggers into roughly L equally-sized sets corresponding to L personality score levels, each one characterized by its mean personality score $\hat{s_i} = \sum_{n=1}^{N_i} s_n/N$, where N_i is the number of videos in the i-th set, and s_n is the personality score corresponding to the n-th video in the set. Then, for each personality trait and each measure of attention, we explore the association between the average personality scores $\hat{s_i}$ and the average level of attention $\hat{a_i}$ obtained for each set $i = 1 \dots L$. For our analysis, we used $L = 20$, as a compromise between the number of videos per set and the

TABLE IV
PEARSON'S CORRELATION COEFFICIENTS BETWEEN VLOGS' ATTENTION MEASURES AND THE PERSONALITY IMPRESSIONS ($^{\dagger}p < .05, ^{*}p < .01, ^{**}p < .001, ^{***}p < .0001$). FOR AGREEABLENESS, THE CORRELATION SCORE RESULTS FROM A SQUARE RELATIONSHIP

Measures	E	C	O	A	ES
# Views	$.93^{***}$	$.61^{*}$	$.78^{***}$	$.70^{*}$	.36
# Times favorited	$.95^{***}$	$.58^{*}$	.82	$-.67^{*}$	$-.09$
# Raters	$.96^{***}$	$.51^{\dagger}$	$.82^{\dagger}$	$.65^{\dagger}$	$-.09$
# Comments	$.97^{***}$	$.61^{*}$	$.78^{***}$	$.60^{\dagger}$	.00
Average rating	.12	.26	$.67^{*}$	$.56^{\dagger}$	.24

number of data points for the analysis.

Figure 5 shows the xyplots of the Big Five personality traits and the attention measure using the number of views. The figure features a linear association between Extraversion, Openness to Experience, and Conscientiousness with attention, suggesting that users scoring higher for these traits receive a higher level of attention from the audiences. Intuitively, it is reasonable to think that vloggers scoring higher in personality traits such as Extraversion or Openness to Experience may be more appealing or interesting to watch, because of the ways in which they create their videos and behave in them (including both the verbal and nonverbal channels). In addition, these type of personalities are more likely to be active and socially involved online as well as offline, and therefore might ultimately be recognized in the vlogger community. Instead, Agreeableness shows a nonlinear association with attention, suggesting that the "pleasant" and "disagreeable" vloggers in the extremes of this personality dimension tend to receive more attention.

We measured the strength of all the possible associations between measures of attention and personality by means of linear fits with the exception of those involving the Agreeableness trait. For Agreeableness, we fit a second-order polynomial with all measures of attention except with the average rating, for which the association was also observed to be linear. Table IV summarizes the correlation coefficient for all the fits (the R^2 values of the second order polynomial fits were converted to correlation coefficients for consistency). The linear and nonlinear associations observed for the number of views in Figure 5 are also measured for the number of times favorited, comments, and raters, whereas the Agreeableness polynomial fit compares in strength to the rest of the fits. The case of average rating is interesting because shows low correlation with Extraversion (and Conscientiousness), suggesting that low and high extraverts are both likely to have high average ratings. On the contrary, the Agreeableness and Openness to Experience trait are linearly associated to the average rating, suggesting that users scoring high in these traits are also obtaining higher average ratings. Thus, the "liking" social function associated to the average rating seems to capture well the meaning of agreeableness, giving lower ratings to disagreeable vloggers and higher ratings to pleasant, likable vloggers.

C. Correlation Analysis between Nonverbal Cues and Personality Impressions

We investigated the individual correlations between automatic nonverbal behavioral cues and personality impressions by means of pair-wise correlations. This is a common approach of research in personality, mostly in social psychology [27], [48] used to explore what aspects of targets observers may have used to infer their judgements (i.e., cue utilization). In our case, this analysis is useful to find out what automatic nonverbal cues are actually capturing such kind of information. Given the highly skewed distribution of some of the automatically computed nonverbal cues, we decided to report all the correlations using Spearman's coefficient, which is better suited for non-normal data (for normal distributed cues, Spearman's coefficient slightly underestimates the true correlation with respect to Pearson's coefficient) [19]. Tables V- VII present the Spearman's correlations between our sets of audio, visual, and multimodal cues and the vloggers' personality impressions. To help the analysis, we also measure the level of cue utilization with the number of significant effects. Previously to our analysis, all the cues were tested for near-zero-variance using a common statistical approach based on the number of unique values per feature and the ratio between the most frequent and the second most frequent value [36]. All cues passed the test. Some nonverbal cues were dropped from the analysis because they showed high correlations with other cues ($r \geq .80$). For example, both median and entropy aggregates of prosodic cues (see Section VI-A2) were dropped because they were found highly correlated with the mean aggregate. The # looking turns was also dropped because of the large correlation with the average length of looking segments.

TABLE V
SPEARMAN'S CORRELATION COEFFICIENTS BETWEEN AUDIO CUES AND THE PERSONALITY IMPRESSIONS ($^{\dagger}p < .05, ^{*}p < .01, ^{**}p < .001, ^{***}p < .0001$).

Speaking Activity	E	C	O	A	ES
Speaking Time	$.18^{***}$	$.27^{***}$	$.13^{*}$	.05	$.12^{**}$
Av Len Speak	$.16^{***}$	$.15^{**}$	$.11^{\dagger}$	.01	.07
# Speech turns	$-.13^{**}$	$-.01$	$-.07$	.03	$-.00$
Cue utilization	3	2	2	0	1
Prosodic cues	E	C	O	A	ES
Voice rate	.02	$.10^{\dagger}$	.05	.09	.04
Av Voice Seg	$-.07$	$-.11^{\dagger}$	$-.07$	$-.09^{\dagger}$	$-.06$
Energy (m)	$.28^{***}$	$-.04$	$.11^{\dagger}$	$-.08$	.01
Energy (m-sd)	$.09^{\dagger}$	$-.13^{\dagger}$	.08	$-.01$	$-.00$
D Energy (m)	$-.08$	.02	$-.08$	.01	$-.00$
D Energy (m-sd)	$.34^{***}$	$-.08$	$.15^{**}$	$-.11^{\dagger}$	$-.04$
Energy (max)	$.32^{***}$	$-.13^{\dagger}$	$.13^{**}$	$-.14^{**}$	$-.08$
Energy (min)	.02	.01	.01	$-.05$	$-.03$
Spec Entropy (m)	.09	$-.08$	$-.02$	$-.04$	$-.01$
Spec Entropy (m-sd)	.05	.01	.06	.05	$-.03$
F0 (m)	$.21^{***}$	$-.08$	.07	$.12^{\dagger}$	$-.05$
F0 (m-sd)	$-.10^{\dagger}$	.01	$-.02$	$-.18^{**}$	.00
F0 (max)	$.20^{***}$	.04	.14	$-.00$	.06
F0 (min)	$.21^{***}$	$-.02$	.07	.01	$-.02$
F0 conf (m)	$.22^{***}$	.02	.07	$.13^{*}$	.02
F0 conf (sd)	$.17^{***}$	.01	.08	$.13^{*}$	.03
F0 BW (m)	$-.26^{***}$	$-.13^{*}$	$-.15^{*}$	.06	$-.02$
F0 BW (m-sd)	.04	$-.00$	.01	.09	.03
F0 Intensity (m)	$.14^{*}$	$.12^{\dagger}$	.09	$-.00$	.07
F0 Intensity (m-sd)	$.21^{***}$	.07	$.14^{*}$	$-.02$	.04
# R0 peaks (m)	$.17^{***}$	$-.09$	$-.01$	$-.04$	$-.04$
# R0 pks (s)	.09	$-.12^{*}$	$-.05$	$-.08$	$-.04$
Loc R0 pks (m)	$.28^{***}$	$-.07$	.06	.03	$-.07$
Loc R0 pks (m-sd)	$-.04$	$-.14^{*}$	$-.07$	$-.14^{*}$	$-.06$
Cue utilization	14	8	5	8	0

m = mean, s = standard deviation, m-sd = mean-scaled standard deviation. All the cues were defined in Section VI-A.

We provide several interesting observations in the light of knowledge of cue utilization for personality impressions and nonverbal behavior. First, we observed that considering audio, visual, and multimodal cues together, Extraversion (cue utilization = 24) and Emotional Stability (cue utilization = 3) are the traits that show the larger and the lower cue utilization respectively. Thus, the finding repeated in a variety of contexts [13], [33], [4], that impressions of Extraversion and Emotional Stability are, among all traits, the ones associated respectively with the largest and lowest amount of informative cues is also supported in the context of vlogging. Second, we noticed that whereas Agreeableness was found to achieve the second largest agreement by far (ICC(1,1) = .27), it only accounted for a small number of significant effects (cue utilization = 10) compared to Conscientiousness (ICC(1,1) = .14, cue utilization = 16) and Openness to Experience (ICC(1,1) = .15, cue utilization = 12). This finding indicates that despite the evidence that observers shared some information during video watching that lead them to agree on their impression of vloggers' Agreeableness, such information is not captured by the features proposed in our datasets. Instead, it seems that the set of cues proposed provide more information related to impressions of Conscientiousness and Openness to Experience. More detailed observations can be found by looking at the different sets of nonverbal cues.

Among audio cues (see Table V), the speaking activity features show significant correlations with all the personality impressions except with Agreeableness. This result concurs with a general finding from social psychology research that reports speaking activity cues among the nonverbal cues with larger number of associations with several social constructs in multiple conversational scenarios [35]. In particular, the positive correlation of speaking time with Extraversion indicates that observers are aware of the general knowledge that associates the extraverts with being talkative [35]. Other correlations are also backed up with related literature. For example, Extraversion judgments are positively correlated with the length of the speech segments and negatively correlated with the number of speaking turns, which agrees with findings that associate Extraversion impressions with high fluency [48].

Some of the effects observed for prosodic cues had also been previously documented in the literature. The positive correlation between Extraversion and both mean/max Energy and mean Pitch is related to Extraversion impressions being associated with people speaking louder [13], [48] and with higher pitch [48]. The positive correlation between Agreeableness and mean Pitch is related to Agreeableness impressions associated with high voice [13]. In addition, the correlations between Conscientiousness impressions and mean-scale standard deviations of Energy, and between both Extraversion and Agreeableness impressions are associated with higher vocal control [35].

The looking patterns (see Table VI) show significant correlations with all the traits except Emotional Stability. In particular, Conscientiousness impressions are associated with vloggers facing the camera longer and more persistently, as measured by the total looking time and the length of looking segments. On the contrary, Extraversion and Openness to Experience judgments are negatively correlated with the length of looking segments. One could argue that the length of looking segments is indicative of gaze avoidance. If that was the case, our results would suggest that Extraverted and Openness to Experience impressions of vloggers are associated with camera avoidance. We found at least one work showing that the association with camera avoidance was negative for most of the personality impressions [13]. One would expect this to be even true also for vlogging, specially when people are voluntarily recording themselves. Thus, it is unclear to what extent the distribution of this nonverbal cue may not be result from body movement or any other behavior instead of gaze avoidance. Clearly, this result needs to be explored in further detail. Regarding pose, the positive correlation of Extraversion, Openness to Experience and Agreeableness with the vertical framing cue suggests that high scores on these traits are associated with vloggers showing the upper body, as opposed to mainly showing the face. Interestingly, recent research in video-based dyadic conversations reported upper body framing to have a significant effect on participants' empathy during interaction [44].

TABLE VI
SPEARMAN'S CORRELATION COEFFICIENTS BETWEEN VISUAL CUES AND THE PERSONALITY IMPRESSIONS
($^{\dagger}p < .05,^{*}\ p < .01,^{**}\ p < .001,^{***}\ p < .0001$).

Looking & Pose	E	C	O	A	ES
Looking time	$-.02$	$.24^{***}$	$-.03$	$.10^{\dagger}$	.09
Av Len Look Seg	$-.13^{*}$	$.23^{***}$	$-.14^{*}$	.07	.07
Proximity to camera	$-.02$	.07	$-.05$	.01	$-.05$
Vertical framing	$.14^{*}$	.00	$.14^{*}$	$.12^{\dagger}$	.08
Cue utilization	2	2	2	2	0
Visual activity	E	C	O	A	ES
wMEI (e)	$.33^{***}$	$-.17^{**}$	$.21^{***}$	$-.01$	$-.03$
wMEI (m)	$.32^{***}$	$-.13^{**}$	$.24^{***}$	.02	$-.00$
wMEI H Cog	.05	$-.04$	$-.01$	$-.06$	.01
wMEI V Cog	$-.04$	$-.05$	$-.08$	$-.03$	$-.06$
Cue utilization	2	2	2	0	0

e = entropy, m= mean. All the cues were defined in Section VI-B1.

The visual activity cues (see Table VI) are among all visual cues, the ones showing the highest correlation values, doing so with Extraversion, as well as with Openness to Experience, and Conscientiousness impressions. As measured by the mean and entropy of wMEI features, high scores of Extraversion and Openness to Experience impressions are associated with high and more diverse visual activity, whereas high scores on Conscientiousness impressions are associated with a vlogging setting involving less and less diverse movement. Apparently, observers seem to share the common knowledge that associates higher levels of activity with enthusiastic, energetic, and dominant people [35]. The same exact findings are reported in related literature, which show that rapid body movements are positively correlated to Extraversion impressions and negatively correlated with Conscientiousness [33], [13].

The multimodal cues also showed a number of significant effects (see Table VII). Large amounts of looking-while-speaking time (L&S) are associated with high scores of Emotional Stability, Extraversion, and Conscientiousness, whereas large amounts of looking-while-not-speaking (L&NS) are also associated with low Extraversion. Note that the total looking

TABLE VII
SPEARMAN'S CORRELATION COEFFICIENTS BETWEEN MULTIMODAL CUES AND THE PERSONALITY IMPRESSIONS
($^{\dagger}p < .05, ^{*}p < .01, ^{**}p < .001, ^{***}p < .0001$).

Multimodal cues	E	C	O	A	ES
L&S	$.14^{**}$	$.29^{***}$	.05	.08	$.12^{*}$
L&NS	$-.16^{**}$	$-.05$	$-.11$	$.06$	$-.07$
L&S/L&NS	$.21^{***}$	$.20^{***}$	$.14^{\dagger}$	$-.02$	$.12^{\dagger}$
Cue utilization	3	2	1	0	2

All the cues were defined in Section VI-C.

time as a single feature (Table VI) did not show any correlation with Extraversion impressions but it does so when combined with speech. Furthermore, the ratio of L&S/L&NS is the multimodal cue showing the largest number of significant correlations, doing so with all the personality impressions, except Agreeableness. In particular, the results regarding Extraversion agree with findings that associate Extraversion impressions with people looking more frequently and with larger glances when speaking (high L&S) [31]. In addition, they concur with findings linking Extraversion to dominant behaviors. For example, in conversational scenarios, higher ratios between looking-while-speaking and looking-while-listening have been found to be associated to impressions of dominance [21].

Summing up, our analysis shows that a number of automatically computed audio, visual, and multimodal nonverbal cues are significantly correlated with vloggers' personality impressions, suggesting that the behaviors measured by these cues may have also been used by the observers. As in related literature, we found that Extraversion impressions showed a significant number of associations to cues from both audio [48], [13] and video [13], whereas Conscientiousness, Agreeableness, and Openness to Experience impressions showed more associations with visual cues [33] . We also found that the low number of correlations of cues with Agreeableness does not explain the high agreement achieved compared to other traits, which motivates the need to look for other cues. For example, some works suggest that facial cues, such as smiling [33], may contain important informations to form impressions about this trait. Finally, the magnitude of the observed effects may compare modestly with effects reported in some social psychology works [4], [48], [33], [13], which report significant correlations between .15 and .70 for a diversity of nonverbal cues. This could be due to different factors. For example, as observed by Yarkoni [59], this could be explained by the fact that effect sizes for statistical significant effects typically vary inversely with the sample size.

D. Automatically Prediction of Personality Impressions using Nonverbal Cues

In this section, we address the task of automatically predicting vloggers' personality impressions. Specifically, we were interested on assessing how individual cue utilization results in prediction performance when using cues together and machine learning techniques for prediction. In addition, we also aimed to evaluate to what extent we can make accurate predictions on the basis of moderately reliable crowdsourced annotations.

We conducted a series of regression tasks (one independent task for each personality trait) targeted to predict the exact score of personality impressions for each of the 442 vloggers in our dataset. For each task, we used a 10-fold cross-validation resampling approach to train and test a Support Vector Machine (SVM) regressor [50]. The 442 instances of our dataset were randomly divided in 10 different folds. At each resample iteration, one fold was used for testing and the rest of 9 folds for training. Each time a model was trained, the parameters were optimized using an inner 10-fold cross validation approach. We evaluated several models with distinct feature sets and different kernels (linear, polynomial, and RBF). Whereas the linear kernel consistently underperformed the RBF and the polynomial kernel, the performance of the RBF and the polynomial kernel was almost the same for all the tasks (only in few cases the RBF provided slightly better performance than the polynomial). Hence, to keep the presentation of the results clear, we decided to only report performance for the RBF kernel. We measured the performance of the automatic predictions using the root mean square error (RMSE) and the coefficient of determination (R^2), as these are the two measures considered in the literature [38], [40]. The baseline regressor is a model that predicts the mean personality score (MPS) of the training data. The RMSE accounts for the average error of the predicted scores:

$$RMSE = \sqrt{\frac{\sum (y_{obs} - y_{pred})^2}{n}}, \tag{1}$$

where y_{obs} and y_{pred} are the observed scores and the predicted scores, respectively, and n is the number of samples. In contrast, the R^2 is measured based on the ratio between the model's absolute error and the baseline-MPS predictor. Formally, it is expressed as:

$$R^2 = 100 \times \left(1 - \frac{\sum (y_{obs} - y_{pred})^2}{\sum (y_{obs} - \bar{y}_{obs})^2}\right), \tag{2}$$

where y_{obs} and $\bar{y}_{obs}$ are the observed scores and their mean, respectively, and y_{pred} are the scores predicted by the model. With this definition, the R^2 can be interpreted as measuring the relative improvement in MSE of the automatic predictor with respect to the baseline-MPS. Note that negative values can be obtained when the automatic predictor fails to outperform the baseline-MPS.

Table VIII shows the performance for the prediction of the Big Five personality impressions averaged over the 10 test folds. To measure significant differences between the models and the baseline, we conducted two-tailed paired t-tests for the RMSE, and two-tailed single t-tests for R^2. Significant improvements are presented in bold. At a first glance, we obtained significant performances compared to the baseline-MPS for three of the Big Five (Extraversion, Conscientiousness, and Openness to Experience), as shown by bold values of R^2 between 7% and 36%. In contrast, as indicated by the low and negative values of R^2, the automatic predictor could not do any better than the baseline for Agreeableness and Openness to Experience traits. Note that for the case of Conscientiousness and Openness to Experience, we see that R^2 values of 7% and above correspond to significant improvements in terms of

TABLE VIII
CROSS-FOLD VALIDATION RESULTS FOR REGRESSION TASK (SIGNIFICANT DIFFERENCES WITH RESPECT TO THE BASELINE SHOWN IN BOLD).

	E		C		O		A		ES	
Feature set	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2
Baseline-MPS	.99		.77		.70		.86		.78	
Audio										
Speech Activity	.96	7	.79	−3	.69	3	.89	−7	.81	−7
Prosody	**.83**	**31**	.76	5	.69	4	.87	−1	.81	−8
Combined	**.83**	**31**	.75	**7**	.68	**6**	.87	−2	.81	−7
Video										
Look & Pose	.98	2	.77	1	.69	4	.90	−8	.80	−5
Visual Activity	.96	**7**	.79	−3	.69	3	.89	−7	.81	−7
Combined	.96	**8**	.76	3	.69	5	.89	−5	.81	−8
Multimodal										
Look & Speech Activity	.98	4	**.74**	**8**	.71	−2	.90	−8	.79	−3
Audio+Video	**.80**	**36**	**.74**	**9**	**.67**	**10**	.87	−1	.80	−5
Audio+Multimodal	**.82**	**31**	**.75**	**8**	.68	5	.87	−2	.81	−7
Video+Multimodal	.96	**8**	.76	3	.69	5	.89	−5	.81	−8
Audio+Video+Multimodal	**.80**	**36**	**.74**	**10**	**.67**	**10**	.87	−2	.80	−4

TABLE IX
CROSS-FOLD VALIDATION RESULTS FOR REGRESSION TASK FOR TWO DIFFERENT HALVES OF THE DATASET (SIGNIFICANT DIFFERENCES WITH RESPECT TO THE BASELINE SHOWN IN BOLD).

	E		C		O		A		ES	
Feature set	N	ICC	N	ICC	N	ICC	N	ICC	N	ICC
High consensus data	203	.90***	308	.84***	253	.77***	257	.84***	331	.81***
Low consensus data	239	.61***	134	.21***	150	.24***	185	.48***	111	.25***
	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2
High consensus data										
Baseline-MPS	1.05		0.88		0.71		0.88		0.86	
Audio+Video+Multimodal	**0.80**	**41**	0.86	**6**	0.71	1	0.9	−4	0.88	−4
Low consensus data										
Baseline-MPS	0.93		0.72		0.69		0.85		0.74	
Audio+Video+Multimodal	**0.78**	**30**	0.69	**8**	0.67	5	0.83	0	0.76	−4

ICC = ICC(1,k)

RMSE, whereas for Extraversion the significance of RMSE is associated to higher values of R^2 due to the largest variance existing in the scores.

Overall, the best performance (up to 36%) is achieved for Extraversion, which is not surprising, given that it is the trait with the largest cue utilization, and the one that achieves the most agreement among observers. Among audio cues, the prosodic cues showed significant performances with respect to the baseline-MPS for both Extraversion (R^2 = 31%) and Conscientiousness (R^2 = 7%), whereas speech activity was only useful to predict Extraversion impressions (R^2 = 7%). For the case of visual cues, visual activity was useful for the prediction of the Extraversion trait (R^2 = 7%), whereas Look and pose cues were useful to predict Conscientiousness (R^2 = 8%). In addition, multimodal cues improved the baseline prediction significantly for Conscientiousness (R^2= 8%). Finally, the combination of audio and visual cues showed improvements with respect to the use of single feature sets alone ($R^2 = 36\%$ for Extraversion, $R^2 = 9\%$ for Conscientiousness, and $R^2 = 10\%$ for Openness), though these differences were not statistically significant compared to using visual cues alone. In addition, the use of multimodal cues did not help to improve the performance compared to audio and video.

We further investigated the effect of the annotations agreement in the prediction of the Big Five personality impressions. We divided the dataset into vlogs that generate consensus among annotators, and those that generate disagreement, by using a measure of ordinal dispersion for likert-scales [54]. For a given target and a set of observations, the consensus estimates the amount of dispersion of the observations with respect to the mean, in a way that it is zero when the same number of observations spread on the extremes of the likert-scale, and one when all the observers agree in one point of the scale. Clearly, grouping samples with higher consensus results in a dataset with higher ICCs, whereas the opposite is true for the samples with low consensus. For each trait, we computed the consensus for each vlogger and we divided the dataset in two parts, with a consensus threshold of .80 (this threshold was kept the same for all the traits). Then, we replicated the regression task for the two different halves of the dataset using all the features.

Table IX shows the size of the samples, the ICC(1,k) achieved, and the regression results for these tasks. For the Extraversion trait, we observed an increase of performance to R^2 = 41% for high consensus data, and a decrease of performance to R^2 = 30% for low consensus data, which indicates that impressions with higher agreement are consistently associated to specific cues with high utilization, in a way that it results easier for the model to automatically learn the personality impressions. However, for the case of Conscientiousness and Openness to Experience, the performance of both halves decreases with respect to the results of Table VIII. This result requires more investigation, as it is unclear to what extent the low performance is due to the size of the sample, the low cue utilization for these traits compared to Extraversion, or both.

Overall, our analysis shows that the task of predicting personality impressions from nonverbal cues is feasible in

vlogging, specially for the Extraversion trait, and that we may need to explore other cues for the rest of the traits, specially for Agreeableness. To contextualize the achieved performances on the prediction of personality, we refer to existing works on the literature described in Section II. For example, Mairesee et al. [40] obtained the best performances for the prediction of Extraversion, Openness, and Emotional Stability with R^2 values of 24%, 18% and 15% respectively based on verbal text content. In [38], Lepri et al. obtained up to R^2 of 22% for self-reported Extraversion based on automatic nonverbal cues.

VIII. Conclusion

We presented a study on personality impressions from brief behavioral slices of conversational video blogs extracted from YouTube. Our approach is based on three novel perspectives. First, we explored the use of crowdsourcing as a way to obtain personality impressions from ordinary people during video-watching, and as a systematic and scalable alternative to collect personality annotations from large-scale data. Second, we investigate the personality impressions in the context of YouTube, and its relation to the social attention received by vloggers. Third, we automatically extracted nonverbal cues from audio and video to describe vloggers' behavior, as opposed to existing approaches that analyze text blog data based on verbal content.

Our analysis of the crowdsourced annotations indicates that the level of agreement achieved by annotators does not differ from those reported in related works, which suggest that, in terms of quality, MTurk may be suitable to collect vlogger personality annotations. In addition, we found the Agreeableness impressions were surprisingly high in vlogging in the light of existing personality research, which suggests that the vlogging setting may be more suitable place to form impressions of this trait, compared to other contexts, such as user profiles, websites, or face-to-face encounters.

On the analysis of personality and social attention, we found evidence that personality impressions might mediate the YouTube vlog watching experience in a way that certain vlogger traits result on audiences watching, commenting, rating, and favoring their videos more. However, these associations were different for Extraversion, Openness to Experience, and Conscientiousness (linear) compared to Agreeableness (u-shape), and were negligible for Emotional Stability. In addition, the average rating showed associations that seem to capture well the meaning of Agreeableness in vlogging.

Regarding cue utilization, our work supports findings that relate Extraversion impressions with audio and visual cues, and Conscientiousness and Agreeableness impressions with more visual information. For the case of Agreeableness, we found that a very small number of the proposed cues captured information that observers could have used in their impressions (i.e., low cue utilization for this trait). In future work, this could be addressed by proposing cues related to smiling or eye gaze, which have shown higher cue utilization in the literature. Finally, our work showed promising results regarding the task of automatic personality impression prediction from nonverbal cues, showing significant performance for Extraversion, Conscientiousness, and Openness to Experience.

Future work can take several directions. For example, we plan to examine alternative ways of aggregating crowdsourced observations to obtain more reliable personality impressions. We may also investigate the accuracy of personality impressions compared to self-reported personality. Finally, we would be interested in the study of the verbal component of vlogging, which should play an important role in social attention and other phenomena.

Acknowledgment

We thank the support of the Swiss National Science Foundation (SNSF) through the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. In addition, we also thank Oya Aran for providing the wMEI visual activity cues, and the YouTube video blogging community.

References

[1] "Oovoo." [Online]. Available: http://www.oovoo.com
[2] "Skyecandy." [Online]. Available: http://www.skyecandy.com
[3] "Videogenie." [Online]. Available: http://www.videogenie.com
[4] N. Ambady, M. Hallahan, and R. Rosenthal, "On judging and being judged accurately in zero-acquaintance situations," *Journal of Personality and Social Psychology*, vol. 69, no. 3, pp. 518–528, 1995.
[5] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.
[6] C. Anderson, O. John, D. Keltner, and A. Kring, "Who attains social status? effects of personality and physical attractiveness in social groups." *Journal of Personality and Social Psychology*, vol. 81, no. 1, p. 116, 2001.
[7] M. Ashton, K. Lee, and S. Paunonen, "What is the central feature of extraversion?: Social attention versus reward sensitivity." *Journal of Personality and Social Psychology*, vol. 83, no. 1, p. 245, 2002.
[8] S. Basu, "Conversational scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2002, supervisor: Pentland, A.S.
[9] L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe, "Please, tell me about yourself: Automatic assessment using short self-presentations," in *Proc. Int. Conf. of Multimodal Interfaces (ICMI-MLMI)*, 2011.
[10] 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. of Weblogs and Social Media (ICWSM)*, 2011.
[11] J.-I. Biel and D. Gatica-Perez, "Voices of vlogging," in *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2010.
[12] J.-I. Biel and G. Gatica-Perez, "Vlogsense: Conversational behavior and social attention in youtube," *ACM Transactions on Multimedia Computing, Communications*, vol. 7, no. 1, pp. 33:1–33:21, 2011.
[13] P. Borkenau and A. Liebler, "Trait inferences: Sources of validity at zero acquaintance," *Journal of Personality and Social Psychology*, no. 62, pp. 645–657, 1992.
[14] A. Brew, D. Greene, and P. Cunningham, "Using crowdsourcing and active learning to track sentiment in online media," in *Proc. of 19th European Conference on Artificial Intelligence (ECAI)*, 2010, pp. 145–150.
[15] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk," *Perspectives on Psychological Science*, vol. 6, no. 1, p. 3, 2011.
[16] J. Burgess and J. Green, *YouTube: Online video and participatory culture*. Polity, Cambridge, UK, 2009.
[17] D. R. Carney, C. R. Colvin, and J. A. Hall, "A thin slice perspective on the accuracy of first impressions," *Journal of Research in Personality*, vol. 41, no. 5, pp. 1054–1072, 2007.
[18] M. Cha, H. Kwak, P. Rodriguez, and Y. Y. Ahn, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proc. the 7th Internet Measurement Conference (IMC)*, October 2007.
[19] G. Corder and D. Foreman, *Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach*. Wiley, 2009.
[20] S. Counts and K. Stecher, "Self-presentation of personality during online profile creation," in *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

[21] J. Dovidio and S. Ellyson, "Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening," *Social Psychology Quarterly*, vol. 45, no. 2, pp. 106–113, 1982.

[22] D. C. Evans, S. D. Gosling, and A. Carroll, "What elements of an online social networking profile predict target-rater agreement in personality impressions?" in *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2008.

[23] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image Vision Computing*, vol. 27, 2009.

[24] A. J. Gill, S. Nowson, and J. Oberlander, "What are they blogging about? Personality, topic and motivation in blogs," in *Proc. Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

[25] S. D. Gosling, S. Gaddis, and S. Vazire, "Personality impressions based on Facebook profiles," in *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2007.

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

[27] S. Gosling, S. Ko, and T. Mannarelli, "A room with a cue: Personality judgments based on offices and bedrooms." *Journal of Research in Personality*, vol. 82, pp. 379–98, 2002.

[28] B. Huberman, "Social attention in the age of the web," *Working together or apart: Promoting the next generation of digital scholarship*, p. 62, 2009.

[29] B. A. Huberman, D. M. Romero, and F. Wu, "Crowdsourcing, attention and productivity," *Journal of Information Science*, vol. 35, no. 6, 2009.

[30] H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Investigating automatic dominance estimation in groups from visual attention and speaking activity," in *Proc. Int. Conf. in Multimodal Interfaces (ICMI)*, 2008.

[31] Y. Iizuka, "Extraversion, introversion and visual interaction," *Perceptual and Motor Skills*, no. 74, pp. 43–50, 1992.

[32] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 17, no. 3, 2009.

[33] D. Kenny, C. Horner, D. Kashy, and L. Chu, "Journal of personality and social psychology," *Consensus at zero acquaintance: replication, behavioral cues, and stability.*, vol. 62, no. 1, pp. 88–97, 1992.

[34] A. Kittur, E. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, 2008, pp. 453–456.

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

[36] M. Kuhn, "Journal of statistical software," *Building Predictive Models in R Using the caret Package.*, vol. 28, no. 1, 2008.

[37] P. G. Lange, "Publicly private and privately public: Social networking on YouTube," *Journal of Computer-Mediated Communication*, vol. 1, no. 13, 2007.

[38] B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro, "Modeling the personality of participants during group interactions," in *Proc. Int. Conf. on User Modeling, Adaptation, and Personalization*, 2009.

[39] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe, "Employing social gaze and speaking activity for automatic determination of the extraversion trait," in *Proc. Int. Conf. on Multimodal Interfaces (ICMI-MLMI)*, 2010.

[40] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–501, 2007.

[41] W. Mason and S. Suri, "A guide to conducting behavioral research on amazon's mechanical turk," *Social Science Research Network Working Paper Series*, 2010.

[42] R. R. McCrae and O. P. John., "An introduction to the five-factor model and its applications," *Journal of Psychology*, vol. 60, pp. 175–215, 1992.

[43] G. Mohammadi, A. Vinciarelli, and M. Mortillaro, "The voice of personality: Mapping nonverbal vocal behavior into trait attributions," in *Proc. ACM Multimedia Workshop on Social Signal Processing*, 2010.

[44] D. Nguyen and J. Canny, "More than face-to-face: Empathy effects of video framing," in *Proc. Int. Conf. on Human factors in Computing Systems (CHI)*, 2009.

[45] S. Nowson and J. Oberlander, "Identifying more bloggers: Towards large scale personality classification of personal weblogs," in *Proc. Int. Conf. on Weblogs and Social Media (ICWSM)*, 2007.

[46] J. Oberlander and S. Nowson, "Whose thumb is it anyway? Classifying author personality from weblog text," in *Proc. the 44th Annual Meeting of the Association for Computational Linguistics*, 2006.

[47] A. S. Pentland, *Honest Signals: How They Shape Our World*, ser. MIT Press Books. The MIT Press, 2008, vol. 1.

[48] K. R. Scherer, "Personality markers in speech," in *Social markers in speech*, K. R. Scherer and H. Giles, Eds. Cambridge: Cambridge University Press, 1979, pp. 147–209.

[49] P. Shrout and J. Fleiss, "Intraclass correlations: Uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, p. 420–428, 1979.

[50] A. Smola and B. Schölkopf, "A Tutorial on Support Vector Regression," Royal Holloway College, University of London, Tech. Rep., 1998.

[51] M. Soleymani and M. Larson, "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus," in *Proc. SIGIR Workshop on Crowdsourcing for Search Evaluation*, 2010.

[52] K. Stecher and S. Counts, "Spontaneous inference of personality traits and effects on memory for online profiles," in *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2008.

[53] F. Steele Jr, D. C. Evans, and R. K. Green, "Is your profile picture worth 1000 words? Photo characteristics associated with personality impression agreement." in *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

[54] W. J. Tastle and M. J. Wierman, "Consensus and dissention: A measure of ordinal dispersion," *International Journal of Approximate Reasoning*, vol. 45, pp. 531–545, August 2007.

[55] Techcrunch, "Shortform Video Platform VYou Reels In $3M," May 22 2011. [Online]. Available: http://techcrunch.com/2011/05/22/shortform-video-platform-vyou-reels-in-3m/

[56] S. Vazire and S. D. Gosling, "e-Perceptions: Personality impressions based on personal websites," *Journal of Research in Personality*, vol. 87, pp. 123–132, 2004.

[57] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, 2002.

[58] M. Wesch, "Youtube and you: Experiences of self-awareness in the context collapse of the recording webcam." *Explorations in Media Ecology*, vol. 8, no. 2, pp. 19–34, 2009.

[59] T. Yarkoni, "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers," *Journal of Research in Personality*, vol. 44, pp. 363–373, 2010.

[60] YouTube Blog, "Thanks, YouTube community, for two BIG gifts on our sixth birthday!" May 25 2011. [Online]. Available: http://youtube-global.blogspot.com/2011/05/thanks-youtube-community-for-two-big.html

Joan-Isaac Biel received the M.Sc. degree in telecommunications engineering from the Technical University of Catalonia (UPC), Barcelona. He carried out his master thesis carried during his stay at the International Computer Science Institute (ICSI), Berkeley. He is currently a Ph.D. student at the Swiss Federal Institute of Technology in Lausanne (EPFL) and research assistant at Idiap Research Institute. His research area of interest is the analysis of sensor and social media data with to the goal to understand human social behavior and communication, as well as to explore how people engage and experiment with multimedia content.

Daniel Gatica-Perez (S'01, M'02) is a Senior Researcher at Idiap Research Institute and Maitre d'Enseignement et de Recherche at the Swiss Federal Institute of Technology in Lausanne (EPFL), where he directs the Social Computing Group. His current work includes methods to understand conversational behavior in social video sites, mobility and communication trends in urban populations of smartphone users, and emerging social phenomena in face-to-face interaction. His research has been supported by the Swiss and US governments, the European Union, and industry. Among other professional activities, he has served as Associate Editor of the IEEE Transactions on Multimedia, Image and Vision Computing, Machine Vision and Applications, and the Journal of Ambient Intelligence and Smart Environments.