Vlogging Over Time: Longitudinal Impressions and Behavior in YouTube

Daniel Gatica-Perez
Idiap Research Institute and EPFL
Switzerland
gatica@idiap.ch

Dairazalia Sanchez-Cortes
Groupe Mutuel
Switzerland
dairazalia@gmail.com

Trinh Minh Tri Do
Trusting Social
Vietnam
minhtrido@gmail.com

Dinesh Babu Jayagopi
IIIT Bangalore
India
jdinesh@iiitb.ac.in

Kazuhiro Otsuka
NTT CSL
Japan
otsuka.kazuhiro@lab.ntt.co.jp

ABSTRACT

YouTube vlogging, as a popular genre of ubiquitous social video, engages people in entertainment, civic, and social activities. Although several aspects of vlogging have been studied in media studies and multimedia analysis, the longitudinal angle of vlogging regarding recognition of personal state and trait impressions from behavior has not been yet analyzed. We present a study using behavioral data of vloggers who posted vlogs on YouTube for a period between three and six years. We use online crowdsourcing to collect a rich set of 21 impression variables for each video, including perceived personality, mood, skills, and expertise. Acoustic and motion features are extracted to characterize basic nonverbal behavior. The analysis shows that only a couple of perceived variables, including perceived expertise and perceived quality of audio and video, display weak temporal patterns. Furthermore, we show that the use of longitudinal data helps to improve the automatic inference of impressions for several of the impression variables.

ACM Classification Keywords

H.5.m. Human-centered computing: Ubiquitous and mobile computing.

Author Keywords

YouTube; social media; vlog; behavior; impressions

1. INTRODUCTION

YouTube is a quintessential case of successful mobile multimedia. YouTube reports to have "over a billion users... and each day those users watch a billion hours of video... where more than half of YouTube views come from mobile devices." [43]. This kind of social video has become relevant not only as a tool for broadcasting entertainment content, but as a communication channel that allows individuals to put themselves on camera and talk, both to self-express and participate in community activities [13, 41]. Conversational social video persists over time, among other factors, thanks to the sustained participation of users. In turn, such sustained video production practices provide opportunities to study behavior displayed on video over time, since many people join social media at an early age and participate online for years, as is the case of YouTube video bloggers (also known as vloggers, a specific type of YouTuber).

Human constructs like personality, friendship, and emotions associated to variations over time have been traditionally studied by psychologists [2, 11, 15, 33]. Most studies in the past have used self-reports collected through questionnaires. In the last decade, longitudinal studies began to exploit online data from social media and other forms of everyday life traces [25]. For example, Roberts et al., [35] presented an analysis of motivation, skills, experience, and participation of software developers and its performance over three years. Archambault et al. [4] captured three years of behavior and attitudes towards social networking (Facebook, LinkedIn, and Twitter). Shakya et al. [38], analyzed connections between well-being and life satisfaction with the use of Facebook over two years. In YouTube, where vloggers often share video over many years with their audiences, work in digital ethnography and human-computer interaction have followed YouTube users over time to understand social video practices from the perspective of digital media literacy (i.e., the skills needed to create and interact effectively with digital media) [22] and of participation in support groups [19].

In this paper, we introduce the longitudinal angle of YouTube conversational video as a novel theme in multimedia analysis. Although YouTube and other sources of online conversational video have been used to automate the inference of personality traits and emotions of vloggers [5, 7, 27, 30], the emphasis has

MUM '18, November 25–28, 2018, Cairo, Egypt

ISBN 978-1-4503-6594-9/18/11…$15.00

DOI: https://doi.org/10.1145/3282894.3282922

so far been on single videos, i.e., on one-sample-per-subject situations. In contrast, we propose to frame the temporal aspect of vlogging as a research subject of its own, and present a longitudinal study of YouTube vloggers' impressions and behavior. We pose three research questions (RQs):

1. What are the levels of inter-observer agreement and correlation that can be obtained for a rich set of perceived attributes of vloggers over time, including personality, mood, and skills?

2. How do impressions about YouTube vloggers change over a substantial period of time (three to six years)?

3. How accurate is the automatic inference of the average social perception of a vlogger over a set of video samples, compared to the inference based on a single video?

These questions are important for two main reasons. First, while longitudinal studies (i.e., studies that follow the same set of subjects for a period of time) have been successfully conducted on text-based social media [4, 38] and psychological research [2, 11, 15, 33], little is known about the case of social video, where people talk on camera about issues they care about, and share this content online. Our research thus contributes to the growing body of behavioral studies in everyday life through emerging technological means, including mobile and ubiquitous devices [25]. The second reason is that, as a rich source of audio-visual data, the longitudinal setting provides the possibility to train inference models based on multiple observations and labels per vlogger, unlike the existing work which uses a single video per vlogger [5, 7, 27, 30], thus enabling better trained models and higher inference performance.

Our paper makes three contributions. As a first contribution, we collected and curated a database of 2376 videos from 99 YouTube users who uploaded conversational videos over a period of three to six years. For each video, we collected crowdsourced impressions from Amazon Mechanical Turk observers for a set of 21 variables including personality, mood, skills, expertise, and technological attributes, resulting in a total of 249,480 individual judgments. Furthermore, we automatically extracted a rich set of nonverbal cues to characterize the behavior of users from audio and video data. As a second contribution, we analyzed the temporal variability of the impression variables, which suggests a weak correlation pattern with respect to time for only a couple of variables, including perceived user expertise and technical audio-visual quality. At the same time, most of the perceived personal traits and states did not show any temporal trend, which also finds support in the psychology literature about the stability of traits like personality, and the transient nature of emotional states. As a third contribution, we show that the use of longitudinal YouTube data helps to improve the automatic inference of impressions from behavioral cues for many of the variables, which could motivate further work to make use of the longitudinal nature of YouTube and other sources of online conversational video to learn better models using multiple samples per user.

The paper is organized as follows. We first review the related literature (Section 2). We then describe the longitudinal YouTube dataset (Section 3), and the procedures followed to collect impressions (Section 4) and to extract behavioral cues (Section 5). We then present and discuss the results for each of our RQs (Sections 6-8). We conclude in Section 9.

2. RELATED WORK

YouTube behavior over time in social sciences and HCI. Seminal ethnographic work on youth and digital literacy in YouTube was conducted by Lange [22], who studied young people and their videos over a two-year period. The study involved the manual analysis of 200 videos and interviews with 150 participants, and is the most detailed recount of social video practices in the social sciences. Digital media literacy is defined as "the ability to access, analyze, and evaluate messages in a variety of forms... including skills such as digital video production that are required to navigate new media environments" ([22], p. 13). One of the theses of [22] is that YouTube young video makers (and vloggers in particular) can indeed acquire "digital skills by choosing preferred media and tackling personally exciting projects" and learn "technical and participatory media skills when they make and distribute videos online" (p. 9). A second thesis is that the acquisition of expertise to create social video is a process that takes time, i.e, "not developed overnight. Media skills are built up through the micro-interactions that individuals have when creating personally interesting media, receiving feedback, and learning to craft the self and broadcast one's message." (p. 10). This work inspires ours with respect to two ideas: examining YouTube video production over time, and studying the perceptual aspect of acquisition of skills by examining impressions of expertise and technical quality provided by viewers.

Other work in HCI has also studied aspects of YouTube over time through online ethnography. As one example, the work in [19] studied 36 YouTube users with HIV, cancer, or diabetes who posted vlogs describing their personal experience with illness. The first and last vlog of each user's channel were selected and manually coded with respect to spoken content, in order to understand how the life experience of users changed over time. Our work also examines a set of vloggers over time, but with other goals and through other means, namely using online crowdsourcing for collection of impressions and automatic audio-visual methods for behavioral extraction, and thus is connected to the literature described next.

Automatic trait analysis of social video users. Research on automatic inference of human traits from conversational social video can be traced back to the work by Biel et al. on recognition of big-5 personality impressions in YouTube video blogs [5, 7, 8, 9]. Other authors proposed to study sentiment analysis of movie reviews shared on online video [28], and the persuasiveness of movie reviewers [30]. There are two common threads to this work. The first one is the collection of impressions using crowdsourcing platforms like Amazon Mechanical Turk, in which workers are asked to complete standard, psychology-validated short instruments [16],[34] to collect trait impressions in a scalable manner. The second common thread of existing work involves the use of audio-visual analysis methods to automatically extract nonverbal features.

The most recent work has shifted towards advanced machine learning methods. Encouraged by the successful use of convolutional neural networks (CNNs) to recognize emotional states from short (1-2 sec) videos [21, 24], approaches for CNN-based multimodal recognition of human traits in social video have recently appeared. They can be categorized between those who use knowledge-driven audio-visual features as input and those that work with raw data input. The work in [29] used a deep neural network for binary classification of persuasiveness of movie reviewers. Knowledge-driven audio and visual features (prosody, facial expressions, etc.) were first extracted. Deep networks with 3-to-10 fully connected layers were learned from these features, and the resulting performance improved over previously reported work [30]. For Big-5 personality impressions, recent approaches using deep networks [44, 42, 17], including recurrent CNNs and residual networks, were applied on a corpus of YouTube video blogs annotated with one question per trait by pairwise rating of videos, showing promising performance on the five traits.

The above work as well as studies on automatic recognition of perceived traits from behavior displayed in face-to-face interactions [23, 3] use only one video sample to make inferences. We contribute by studying longitudinal aspects of YouTube behavior in the context of 21 impression labels available for over 20 videos per user, which extends what has been attempted in previous work. Finally, note that our goal is in using well-known behavioral features as a testbed for this first study, rather than targeting the best possible performance.

3. LONGITUDINAL YOUTUBE DATASET

To investigate longitudinal phenomena in YouTube, we need data generated by the same people over multiple years. The dataset collected by Biel and Gatica-Perez provided an initial set of vloggers as the basis for a longitudinal collection [7]. These are 442 users (208 male and 234 female) appearing in videos where only one person talks on camera. For this set of vloggers, we downloaded all videos posted over a five-year period, between January 2006 and March 2011. This resulted in 37,000 videos from 421 vloggers. We found variations in the number of videos uploaded per user, which reflects typical user trends in social media (e.g., some users uploaded over 250 videos in the studied period; while many others uploaded 50 videos or less).

We used the concept of conversational video segment as the basis to select users and videos for our study. Conversational segments are defined as continuous video shots in which (1) speech is automatically detected, and (2) the vlogger's face is also automatically detected in at least 60% of the video frames. In other words, we are interested in segments where the vlogger is alone on camera and speaking (instead of doing other activities that can occur in vlogs like dancing, showing artifacts on camera, or adding animations). Furthermore, we used 30-second conversational segments as the video unit for our experiments. Previous behavioral studies on first impressions show that 30 seconds represent a reasonable duration for perception tasks of human traits [1]. With these criteria, we retained those users who had at least twenty videos throughout the entire period of study, each containing at least one 30-second conversational segment, resulting in 197 users. We assumed these users to be reasonably active users. We then made a manual check and eliminated users who sometimes uploaded videos featuring other people, resulting in 177 users.

Next, we chose vloggers who were longitudinally active and sampled videos for our study. We assumed vloggers as longitudinally active if they had uploaded data for three or more years. This resulted in 99 users. The exact distribution of vloggers with respect to years of data with this condition was as follows: 45 vloggers had three years of data, 39 had four years, 12 had five years, and three had six years. Furthermore, so as to capture changes over time, we sampled a fixed number of 24 videos for every user (12 from their early period and 12 from their late period.) Note that given the differences in temporal persistence of each user (3-6 years), this corresponds to a non-uniform sampling over time. After these preprocessing steps, our final longitudinal dataset consists of 2376 videos (99 vloggers, 24 videos per vlogger), and features 61 males and 38 females.

4. CROWDSOURCED COLLECTION OF IMPRESSIONS

To study how vloggers are perceived, we designed a 21-item questionnaire about impressions with respect to four major factors - technical quality, traits, mood, and skills. In this section we first describe the questionnaire, and then describe the implementation of a process to collect impressions via crowdsourcing.

Questionnaire. A total of 21 impression attributes were used as listed in Table 1, categorized into technical quality, traits, mood, and skill factors.

Technical Quality. The first set of items relates to the perceived technical quality of the video. It is known that the quality of the video or audio tracks can affect the watching experience and so impact the overall perception of the expertise of the video maker [22].

Traits. Extraversion and agreeableness, two of the Big Five traits, were included with two questions each, chosen from the Ten-Item Personality Inventory (TIPI) [16]. We dropped the other three Big-Five traits, as work by Biel and Gatica-Perez [7] reported that compared to Extraversion and Agreeableness, the reliability measures for Conscientiousness, Openness to Experience, and Neuroticism were lower. This choice also allowed us to collect other new impression labels. Specifically, we included physical appearance (pretty), and other traits likely to appear in the data (intelligent, funny, quiet, critical). Furthermore, we adapted five items from questionnaires related to the narcissism trait [14], [18], as this trait has been previously reported to be elicited in social media [12].

Mood. We included a few mood-related items which had shown high reliability in previous work [7], including happy, angry, and bored.

Skills. We included three items to collect impressions of persuasiveness, ingenuity, and expertise, which are personal factors that affect viewers' responses in social video [30, 22].

Table 1: Categorized list of 21 impression attributes.

Technology	Trait	Mood	Skills
image quality	extraversion	happy	persuasive
audio quality	agreeableness	angry	ingenious
edition quality	pretty	bored	expertise
	intelligent		
	funny		
	quiet		
	critical		
	narcissism		
	attention centred		
	special		
	storyteller		
	manipulative		

Items related to technical quality and expertise were labeled on a five-point scale. All the remaining items were labeled on a seven-point Likert scale.

Crowdsourcing impressions via Mechanical Turk. We collected five impressions per video for each item in the questionnaire. The number of workers per video was based on previous literature [6] that shows that this number produces annotations of good quality in terms of reliability. With this choice, we ended up with 5*2376 = 11880 Human Intelligence Tasks (HITs), and a total of 11880*21 = 249,480 individual judgments. A HIT took two minutes to complete. We divided the HITs into batches so that the MTurk workers could not see the same vlogger again. The country of residence of the workers was selected to be USA, to reduce variations due to cross-cultural factors. We employed 532 annotators in total. Our design of the HIT made sure that the 30-second video was played before the questionnaire appeared. Every section of the questionnaire (quality, traits, etc.) was verified so that only one section was active before the next subsection appeared. The last section had verification questions asking about the gender of the vlogger and whether they wore glasses. Finally, we also dynamically spotted and eliminated data from potential spammers, discarding their contributions. The HITs were also restricted to workers with acceptance rates of 95% or higher as MTurk workers.

5. NONVERBAL FEATURE EXTRACTION

Table 2 lists the audio and visual nonverbal features extracted from each vlog. They are summarized in this section. Our intention was not to be exhaustive regarding the possible list of extracted cues, but rather to study features that have been useful in previous work and that are interpretable. This is also the reason why we did not include advanced features derived from deep learning, which are not necessarily interpretable.

Audio features

For the audio channel, we first computed several acoustic features grouped into intensity, formants, and pitch using Praat [10]. We computed formants where each frame contains frequency (F0 to F3) and bandwidth (B0 to B3) estimated using a window size of 0.02 seconds. Pitch was estimated per frame using an autocorrelation method, using a time window of 0.01 seconds (i.e. 100 pitch values per second). The intensity of the audio channel (in dB) was also estimated at frame level with a window of 0.01 seconds. From the described features, we estimated several statistics including mean, median, standard deviation, variation (sd/mean), entropy, maximum, and minimum.

In addition, we estimated another set of audio features using the MIT toolbox [32]. The features include confidence in formant frequency, value of largest autocorrelation peak, location of largest autocorrelation peak, number of autocorrelation peaks, averaged voiced segments, averaged length of voice segments, voice rate per second, number of turns, and speaking time. The list of extracted features is summarized in Table 2.

Table 2: Nonverbal features from audio (first to third block) and video (last block). BW: Bandwidth, Freq: Frequency.

Feature	Feature ID	Feature	Feature ID
Max Intensity in frame	IntensityF	Mean of Intensity curve in dB	Intensity
Freq Formant0	F0	BW Formant0	BW0
Freq Formant1	F1	BW Formant1	BW1
Freq Formant2	F2	BW Formant2	BW2
Freq Formant3	F3	BW Formant3	BW3
Number of Pitch candidates per frame	nPcands	Candidate's periodicity	PStrength
Candidate's freq Hz	IntensityP		
Computed: mean, median, standard deviation (sd), variation=sd/mean (cov), entropy, max,min			
Energy in frame	energy	Time derivative of energy in frame	d.energy
Confidence in Pitch	conf.pitch	Num of auto correlated peaks	num.apeak
Autocorrelation peak	apeak		
Location of auto correlation peak	loc.apeak		
Computed: mean, standard deviation (sd)			
voiced segments	avg.voiced.seg	length of segments	avg.len.seg
speaking time	time.speaking	number of turns	num.turns
voice rate	voice.rate	voice rate seconds	voice.rate.sec
weighted Motion Energy Image	wmei		
Computed: mean, median, standard deviation (sd), variation=sd/mean (cov), entropy, max,min			

Visual features

We estimated the weighted motion energy image (wMEI), which has shown to be an informative nonverbal cue for personality traits [7], mood [37] and leadership [36], as it captures the overall visual dynamism of a vlogger. A wMEI is a gray scale image that quantifies motion in a video, where brighter pixels correspond to areas with higher motion. We normalized the outputs per vlog and we estimated the mean, median, standard deviation, entropy, maximum, and minimum.

6. ANALYSIS OF IMPRESSIONS (RQ1)

We now present an analysis of the collected impressions in terms of descriptive statistics, reliability, and correlation.

Descriptive Statistics

Table 3 presents descriptive statistics of the collected impressions at the level of videos (N=2376). The impression scores of each video for each vlogger were aggregated across the corresponding five annotations by computing the mean. Only three traits (extraversion, agreeableness, intelligent) and one mood (happy) have mean values above 4 (the mid-point of the seven point scale). Comparing with previous work, the statistics of extraversion and agreeableness in this dataset are

comparable with those reported by Biel and Gatica-Perez [7]. The mean (and std) values of extraversion and agreeableness in our dataset are 4.99 (1.01) and 4.67 (0.72), as compared to 4.61 (1.00) and 4.68 (0.87), respectively, reported in [7]. Furthermore, mood and attractiveness can also be compared with previous results. The mean (and standard deviation) of our dataset are for happy: 4.35 (0.94); angry: 2.02 (0.83); bored: 2.46 (0.85); and pretty: 3.91 (0.99). This can be compared to the corresponding values reported in [6], namely happy: 4.32 (1.18); angry: 2.15 (1.10); bored: 2.41 (1.04); and beautiful: 4.41 (1.02).

Table 3: Descriptive statistics of the attributes.

Attribute	mean	sd	min	max	ICC
image quality	3.23	0.63	1	5	0.64
audio quality	3.4	0.60	1.2	5	0.50
edition quality	2.71	0.68	1	5	0.40
extraversion	4.99	1.01	1.3	7	0.69
agreeableness	4.67	0.72	1.3	6.7	0.43
pretty	3.91	0.99	1.2	6.8	0.59
intelligent	4.28	0.70	1.6	6.8	0.52
funny	3.17	0.97	1	6.4	0.51
quiet	2.69	1.07	1	6.2	0.67
critical	2.48	0.91	1	6.6	0.52
narcissist	3.13	0.99	1	6.6	0.35
attention centric	3.51	0.67	1.4	5	0.52
special	3.37	0.62	1.4	5	0.38
storyteller	3.50	0.60	1.6	5	0.35
manipulative	2.47	0.60	1	4.6	0.23
happy	4.35	0.94	1	7	0.55
angry	2.02	0.83	1	6.6	0.69
bored	2.46	0.85	1	6.8	0.45
persuasive	3.64	0.80	1.2	6.4	0.23
ingenious	3.40	0.86	1	6.4	0.37
expertise	2.97	0.62	1	5	0.57

Reliability Analysis

We computed the reliability of the impression attributes using ICC (intraclass correlation coefficient). This is commonly used to measure how well annotators agree on a certain judgment [40]. We used the ICC(1,k) estimate, where each target video is rated by a different judge and the judges are selected at random, with k raters from a set of K possible raters. In our experiments $k = 5$ and $K = 532$ (the number of crowdworkers). Table 3 also shows the ICC scores for all the variables.

Among the technology factors, image quality had the highest ICC score (0.64) followed by audio quality (0.50). Regarding the trait factors, extraversion had high ICC score (0.69). Annotators also have good agreement on physical attractiveness (0.59) and quiet (0.67). Our ICC scores on extraversion and attractiveness are close to Biel et al. [6] who reported 0.76 for extraversion and 0.69 for beautiful. Other traits having agreement above 0.5 include intelligent, critical, attention centric, and funny. The rest of traits have lower ICCs, including four of the five dimensions related to narcissism, which suggests that this construct is not evident to rate in the studied setting.

For the mood factors, annotators agree on happy (0.55) and angry (0.69), and less on bored (0.45). Our ICC scores on mood are comparable with the ones reported in [37], where their reported values are happy (0.76), angry (0.67) and bored (0.45). Finally, regarding the skill factors, expertise has the highest agreement (0.57). In contrast, the agreement on ingenuity and persuasiveness were both below 0.4. In the rest of the paper, we will continue the analysis for the 21 variables for purposes of completeness, with the clear understanding that several variables have low ICC (six of the 21 variables have ICC below 0.4).

Correlation Analysis

For each pair of variables, we compute the Pearson correlation coefficient. This is shown in Table 4. We now discuss some of the observed correlations (with absolute value above 0.3).

Regarding Big-Five traits, extraverted vloggers were perceived to be funny (r=0.56), and neither quiet (r=-0.73) not bored (r=-0.59). The literature on extraversion establishes that extraverts are generally cheerful and talkative [20]. Therefore, being perceived as funny and neither quiet nor bored appears to be intuitive. Moreover, for the narcissism trait, we observe a negative correlation between narcissist and agreeableness (-0.39), followed by positive correlations with angry (0.34), critical (0.33), and extraversion (0.31). Attention centric is correlated with extraversion (0.65), followed by a positive correlation with narcissism (0.56), and it is negatively correlated with quiet (-0.56). Attention centric, special, storyteller, and manipulative are correlated with one another (r above 0.62). The observed correlations with extraversion and agreeableness match previous findings between narcissism and the Big-Five traits [31], in which narcissism was negatively correlated with agreeableness (r=-0.36), and positively correlated with extraversion (r=0.41).

According to Table 4, agreeable vloggers are perceived as happier (0.61) and less angry (-0.72). Literature on agreeableness states that agreeable people tend to be warm, pleasant and kind. Appearing to be happy could contribute to also being perceived as pleasant and warm. This said, there is some literature in psychology that does not support this correlation between happiness or joy with agreeableness [39], in which the correlation is found to be close to zero.

Regarding the skills factor, the perception of being ingenious has correlation to the perception of being intelligent (0.5) and funny (0.68). The perception of being persuasive seems to relate to perceived intelligence (0.60) and perceived ingenuity (0.65). Some related work has suggested that fast-talking people can be perceived as more credible, intelligent, and persuasive [26]. Note however that these results have to be taken with caution given the low ICC obtained for both persuasion and ingenious.

Finally, perceived expertise (the skill in making social video) correlates with the technological factors: quality of the images (0.74), audio (0.72), and edition (0.68). Furthermore, perceived expertise is also linked to extraversion (0.45), funny (0.44), and ingenious (0.54). Taken together, these results could suggest a link to the work on YouTube digital literacy in [22], as "when creators post a video, they often transmit - whether intentionally or not - important information about their technical ability" ([22], p. 21).

7. MODELING IMPRESSION CHANGES OVER TIME (RQ2)

We now proceed to quantify possible changes of vlogger impressions over time. First, we present a principled approach

Table 4: Correlation of all impression variables (N=2376). Entries marked with $(+)$ correspond to $p < 0.01$; otherwise $p << 0.01$.

	Attributes	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	image																				
2	audio	.76																			
3	edition	.61	.57																		
4	extraversion	.34	.38	.31																	
5	agreeableness	.08	.08	$.03^{+}$	-.06																
6	pretty	.30	.25	.18	.23	.32															
7	intelligent	.36	.35	.29	.14	.29	.37														
8	funny	.36	.36	.39	.56	.16	.32	.30													
9	quiet	-.28	-.33	-.28	-.73	.23	-.10	$.02^{+}$	-.40												
10	critical	$-.04^{+}$	$-.02^{+}$	$.01^{+}$	$.03^{+}$	-.45	-.18	$-.04^{+}$	$-.02^{+}$	-.09											
11	narcissist	.09	.07	.06	.31	-.39	.09	-.13	.14	-.21	.33										
12	attentioncentric	.25	.25	.24	.65	-.20	.25	$-.02^{+}$	.42	-.56	.20	.56									
13	special	.21	.22	.22	.54	-.19	.22	$-.02^{+}$	.32	-.46	$.21^{+}$	.54	.85								
14	storyteller	.25	.30	.24	.51	-.13	.19	.17	.34	-.41	$.21^{+}$	.43	.71	.72							
15	manipulative	.21	.21	.24	.41	-.29	.20	.07	.26	-.32	.38	.55	.63	.64	.62						
16	happy	.25	.25	.22	.36	.62	.41	.34	.49	-.18	-.45	-.06	.20	.16	.19	.05					
17	angry	-.03	-.04	$-.01^{+}$	$-.01^{+}$	-.70	-.14	-.09	-.09	-.02	.60	.34	.12	$.13^{+}$	.11	.31	-.56				
18	bored	-.29	-.34	-.24	-.59	-.21	-.23	-.24	-.33	.53	.23	$.07^{+}$	-.29	-.24	-.27	-.10	-.34	.25			
19	persuasive	.37	.40	.37	.39	.21	.35	.60	.52	-.24	$.01^{+}$	$.02^{+}$	.24	.23	.37	.31	.42	$-.01^{+}$	-.35		
20	ingenious	.44	.44	.54	.43	.19	.32	.50	.68	-.31	-.04	.09	.30	.26	.33	.25	.44	-.08	-.33	.65	
21	expert	.74	.72	.68	.45	$.01^{+}$	.22	.35	.44	-.38	.02	.13	.32	.28	.35	.28	.25	$.01^{+}$	-.35	.44	.54

to study the relationship between these vlog attributes and the vlog upload time. Then, we apply this approach to characterize the temporal patterns of impressions.

An approach to identify temporal patterns

We consider impressions as random variables and study the relationship between these variables with the vlog upload time, which is cast as another random variable. Upload time is a proxy for the actual vlog creation time.

One simple method is to estimate the Pearson correlation coefficient between the variable of interest and the upload time for each user. A trend for the population of vloggers could then be observed from the statistics of the correlation values.

Another method to identify temporal patterns is to introduce explicit temporal models, and validate these models on the collected data. This approach allows to examine several assumptions about the changes of impression over time. We study three models: (1) an invariant model that assumes that the variable of interest does not change over time; (2) a generic linear model assuming that the variable of interest depends linearly on time and that all users have the same trend; and (3) a user-specific linear model that assumes that the relationship between time and the variable of interest varies depending on the user (e.g., a variable increases for some users, but decreases for other users). Formally, let $f(u,t)$ be the observed value for user u and time t. We introduce several approximation functions $g(u,t)$ that reflect different assumptions about the data. Note that we omit the identifier of each impression for simplicity, but the models are applied to all impression variables.

Invariant model. The first model assumes that the variable of interest is invariant for all videos coming from the same user:

$$g^{invar}(u,t) = c_u,$$

where c_u is a user-dependent constant. The performance of this model reflects how consistent the random variable is, where a high performance in terms of data fitting indicates that the variable is highly consistent.

Generic temporal model. This model assumes that the variable of interest has a linear relationship with time and that the relationship is generic for all users:

$$g^{gFit}(u,t) = a \cdot t + c_u,$$

where a represents the (positive or negative) slope of the regression line. A positive value of a indicates that the variable generally increases over time, and decreases for a negative value.

User-specific temporal model. The last model assumes that the variable of interest has a linear relationship with time, but this relationship depends on the user:

$$g^{uFit}(u,t) = a_u \cdot t + c_u,$$

where a_u is the slope of the regression line, which could be different for different users.

Validation of temporal models. The models are evaluated on the full vlog dataset. For this analysis, we do not use cross validation. The whole dataset is used for training and testing the model. Regarding the evaluation measure, we use the standard R^2 measure.

Correlation with respect to time

Pearson correlations between perceived impressions and time ($N_t = 24$ for each variable) were calculated for each user. Average, median, standard deviation, maximum, and minimum correlations are reported in Table 5. For each user, the time was shifted based on the user's first vlog in the dataset, which corresponds to a user-specific start time. Further videos are realigned, and timestamps are transformed into epoch units. We also display the number of users for whom the corresponding correlation has a p-value < 0.05, according to its sign (positive or negative correlation). The values in these two columns thus range between 0 and 99. Finally, in the last column, we show the median of the correlation computed for those users

Table 5: Statistics of Pearson correlations per user with respect to time. The last three columns show the number of users for which there are negative (-) and positive (+) correlations with time (p<0.05), and the median of the corresponding correlation (only when the number of users is larger than 25).

Attributes	mean	med	sd	max	min	corr ($p < 0.05$)		
						-	+	med
image	0.36	0.44	0.27	0.80	-0.36	0	55	0.54
audio	0.31	0.35	0.26	0.73	-0.41	1	38	0.56
edition	0.19	0.22	0.27	0.80	-0.48	1	27	0.46
extraversion	0.13	0.15	0.30	0.68	-0.68	7	18	≀
agreeableness	0.01	0.02	0.25	0.60	-0.52	2	7	≀
pretty	0.03	0.05	0.26	0.77	-0.49	6	7	≀
intelligent	0.09	0.09	0.20	0.56	-0.31	0	9	≀
funny	0.16	0.16	0.23	0.67	-0.44	1	14	≀
quiet	-0.06	-0.08	0.29	0.71	-0.71	0	15	≀
critical	0.01	0.05	0.26	0.43	-0.60	7	2	≀
narcissist	-0.01	-0.02	0.23	0.51	-0.44	3	3	≀
attentioncentric	0.01	0.02	0.25	0.46	-0.67	6	4	≀
special	-0.01	0.00	0.25	0.58	-0.59	7	3	≀
storyteller	0.03	0.02	0.25	0.55	-0.69	4	6	≀
manipulative	0.06	0.08	0.25	0.58	-0.65	3	5	≀
happy	0.08	0.09	0.23	0.64	-0.54	4	7	≀
angry	0.04	0.05	0.24	0.54	-0.44	2	6	≀
bored	-0.05	-0.10	0.27	0.61	-0.66	9	6	≀
persuasive	0.15	0.15	0.24	0.72	-0.32	0	7	≀
ingenious	0.16	0.18	0.24	0.68	-0.37	0	15	≀
expertise	0.33	0.39	0.27	0.81	-0.42	1	44	0.54

for whom p<0.05, and only for those variables for which at least 25 of the users have correlations with p<0.05, which corresponds to a minimum of 25% of the users.

As we can see, the mean and median correlation between image quality and time are 0.36 and 0.44. Moreover, 55 users have positive correlations with p<0.05, with median equal to 0.54. For audio quality, the mean and median correlation values are 0.31 and 0.35, and 38 users have positive correlations with p<0.05, with median equal to 0.56. This suggests that for the set of users under study, the perceived technological means of producing vlogs show a weak trend to increase over time. This is not surprising given the fact that audio-visual technology has indeed made consistent progress in terms of image resolution and audio quality. We observe a similar trend for perceived expertise: mean and median correlations of 0.33 and 0.39; and median of 0.54 from the 44 users with positive correlations and p<0.05. In contrast, the rest of the perceived attributes reported in Table 5 have null or marginal correlation with respect to time as a population (although positive and negative correlations exist for a few specific users).

Changes of impressions over time

We now analyze the results of our temporal models to validate the three assumptions regarding temporal changes on the analysis of impressions about vloggers across their videos. The variability of the impressions about vloggers is quantified by the R^2 value of the Invariant model in Table 6. Physical appearance is, not surprisingly, the most consistent variable. However, the R^2 value of 0.52 indicates that the perception on beauty still varies across videos. Other variables in the top five most consistent impressions are extraversion (R^2=0.4), quiet (R^2=0.37), expertise (R^2=0.33), and attention centric (R^2=0.3). On the other hand, narcissism, angry, and intelligent are the least consistent variables. The results of the Generic model

Table 6: Data fitting performance (R^2) of models for the set of impression attributes, sorted for the invariant model (Invar). The gFit and uFit bars correspond to the generic temporal model and the user-specific temporal model, respectively.

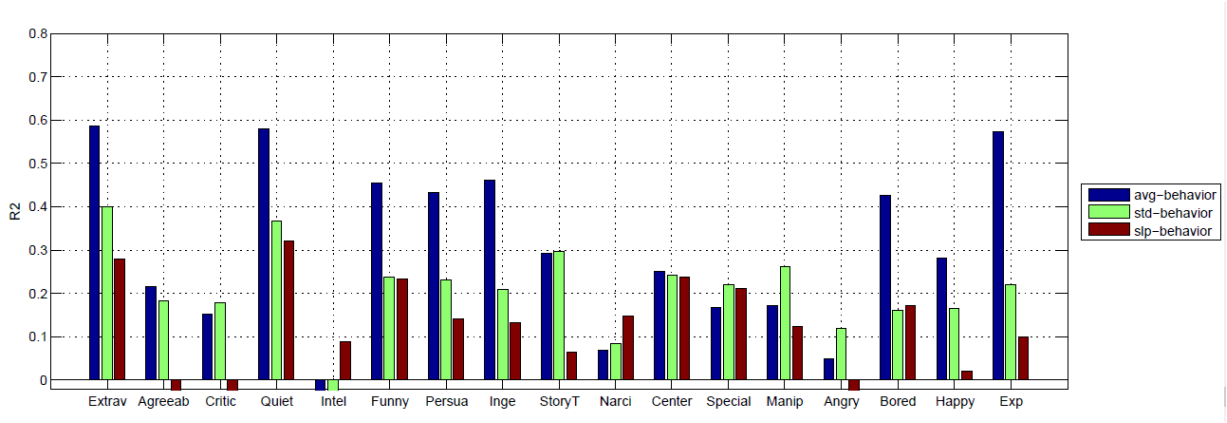
Attributes	R^2 Invar.	R^2 gFit	R^2 uFit
pretty	0.53	0.53	0.56
extraversion	0.40	0.42	0.47
quiet	0.37	0.37	0.43
expertise	0.33	0.41	0.47
attentioncentric	0.31	0.31	0.35
image	0.29	0.39	0.44
edition	0.25	0.28	0.35
audio	0.25	0.33	0.39
funny	0.24	0.26	0.31
ingenious	0.24	0.26	0.31
agreeableness	0.23	0.23	0.28
special	0.23	0.23	0.28
happy	0.22	0.23	0.27
bored	0.22	0.22	0.29
critical	0.21	0.21	0.27
storyteller	0.21	0.21	0.26
manipulative	0.20	0.20	0.25
persuasive	0.19	0.21	0.26
narcissist	0.16	0.16	0.20
angry	0.12	0.12	0.19
intelligent	0.11	0.12	0.17

(gFit) and User-specific model (uFit) share a similar trend for some variables: the Generic model improves over the Invariant model on Expertise, Image Quality, and Audio Quality (as suggested by the correlation analysis). Furthermore, the User-specific model further improves over the Generic model in terms of R^2. One can see that the largest improvements of the User-specific model over the Invariant one are for image quality, audio quality, and expertise (R^2 gains of 0.14 to 0.15 for each of these variables).

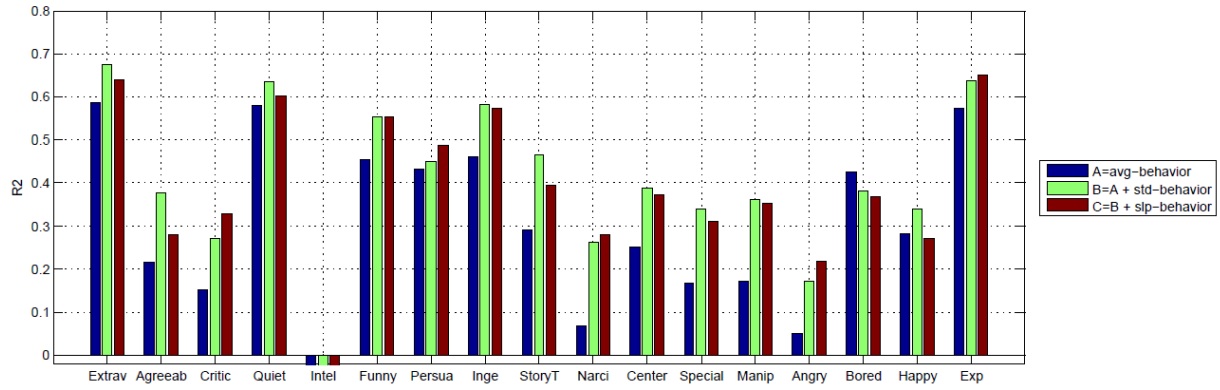
The results in Table 6 suggest that as the time information cannot fully explain the existing variabilities in terms of data fitting, other factors that affect the variability on impressions might come into play, including the intrinsic uncertainty of the impression variables, within-annotator variability, and content-specific factors (e.g. the topic discussed in the vlog). Future work could investigate in more depth these possible sources of variability.

8. AUTOMATIC INFERENCE OF IMPRESSIONS (RQ3)

We now present results on automatic inference of vlogger impressions. While the inference of perceived traits from vlogs has been addressed in the past, the novelty of our analysis is the definition of a new task in which the goal is to infer a given perceived trait based on *a collection of videos* instead of a single one. This setting raises several questions. First, what are good features that can be extracted from a collection of videos? For example, can behavioral changes can help improving the accuracy of the system?. Second, we expect that the difficulty of inferring the overall impressions on a set of videos would be different than the difficulty of making inferences on single videos. A collection of videos contains more behavioral information than a single video, which could help improve the accuracy of automatic inference.



(a) Individual setting.



(b) Incremental setting.

Figure 1: Automatic inference of perceived traits and other attributes at the user level.

Feature selection

From Section 5, there are 142 extracted features in total, but not all of them might have a strong relationship with the impression variables in this study. Furthermore, some features may contain redundant information. For these reasons, we performed a feature selection step.

We used 10-fold cross-validation (CV) for feature selection (FS) for the multi-label prediction task of the 21 impression variables. For each test fold of the CV loop, we train on the 9 remaining folds, thus training and testing are well separated. First, we compute the correlation matrix between the 142 features and the 21 impression variables. Then, we rank the set of features by the average correlation coefficient over the 21 variables, before employing a forward feature selection method. Starting with an empty working set, the method iterates over the ordered set of training features and adds a feature to the working set if it helps improving the cross-validation performance of predicting the 21 impression variables. To evaluate performance of a given set of features, we use linear regression and compute the root mean square errors (RMSEs) for each of the 21 impression attributes. The overall prediction performance is the average value of all RMSEs. At the end of the feature selection step, we obtain a core set of 55 features.

This feature selection method works for data having multiple output variables, unlike the standard setting of considering only one output variable at a time. We experimentally found that this approach is less prone to overfitting. This can be explained by the fact that the evaluation criterion, defined on multiple output variables, is more robust than the one based on a single variable.

Automatic inference

For the inference of perceived traits at the user level, our dataset consists of $N = 99$ users, each of them has 24 videos. The output variables are the averages of impression scores for each of the impression attributes. Furthermore, for each user we compute three feature groups: the averages of behavioral features, their standard deviations, and their slopes with respect to time (i.e., the rate at which each feature increases or decreases over time which indicates the trends of behavioral change). We use linear regression for the experiments, and the results are computed in a 10-fold cross-validation setting.

To estimate the importance of each feature group, we perform two series of experiments, namely an individual setting and an incremental setting. In the first setting, we evaluate the performance when using each feature group separately. The

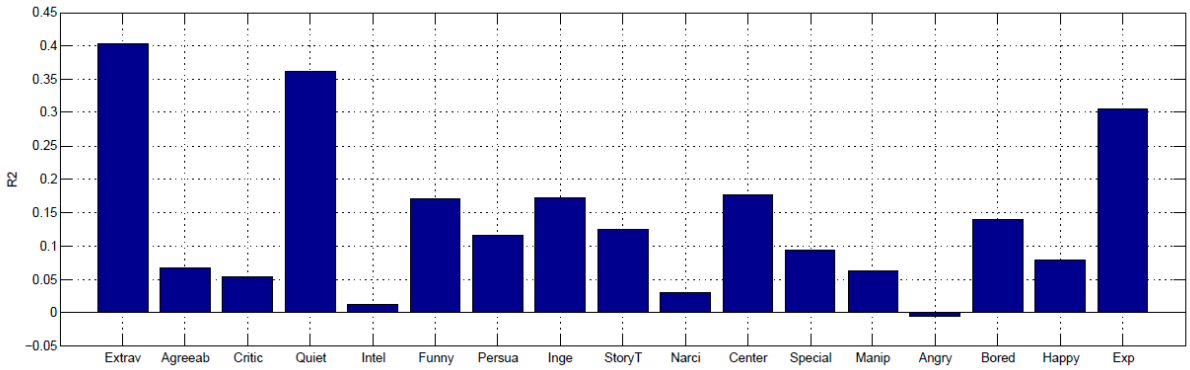


Figure 2: Automatic inference of perceived traits and other attributes at the video level.

results in Figure 1a show that the average of behavioral features are the best features for most of the impression variables. Extraversion, Quiet, Funny, Bored, Persuasive, Ingenious, and Expert are the variables with highest R^2 scores, ranging from 0.4 to 0.6 using only the first feature group (average behavior). The two other groups of features (standard deviation and slope of behavior) give less accurate inference, but the results still show connections between these features and the traits. The standard deviation group contains the best features for Angry, StoryTeller, and Manipulative. In the second setting of experiments, we evaluate the performance for multiple groups of features in Figure 1b. Looking at the impressions for which the averages of behavior are good features (i.e., Extraversion, Quiet, Funny, Bored, Persuasive, Ingenious, and Expert), we find that the two other feature groups could improve the R^2 score of the regression model in all cases except for Bored. The combination of the average behavior group and the standard deviation group results in the best performance in general. Critical and Angry are the two exceptions, for which the combination of all three groups of features is most effective.

Finally, to make comparisons between the above results with the inference results at the video level, we run linear regression on a dataset of $N = 2376$ videos and report 10-fold cross validation results in Figure 2 (variables are displayed in the same order). Our result for inferring Extraversion at the video level is $R^2 = 0.4$, which is generally comparable with the result of $R^2 = 0.36$ reported in previous work [7], but note that a strict comparison is not possible as the datasets are different (2376 videos in our case vs. 442 videos in [7]). On the other hand, comparing the results of automatic inference at the user level and at the video level, we find that inference at the user level seems more reliable than at the video level. For example, the R^2 performance for Extraversion increased from 0.4 at the video level to 0.68 at the user level. A similar trend is seen for other variables. Also note that some variables such as Critical are almost unpredictable at the video level (partly due to the low ICC discussed in Section 6), but achieve better performance at the user level. This suggests that when multiple videos of a vlogger are available, basic aggregates of behavioral cues can result in inference improvements. This is important as it suggests the advantages of thinking in longitudinal terms with respect to vlogger analysis.

9. CONCLUSION

In this paper, we presented a longitudinal study of YouTube vloggers' impressions and behavior as a novel theme in social media. We conclude by summarizing the answers to the three RQs we posed.

RQ1 inquired about the levels of inter-observer agreement and correlation for a set of 21 perceived vlogger attributes, including personality, mood, and skills, collected using crowdsourcing. The ICCs showed inter-observer agreement values above 0.5 for 12 of the studied variables. A correlation analysis further revealed multiple connections, several of which match what has been reported in previous literature, but making use of five times more data.

RQ2 asked whether impressions made about vloggers changed over time. We used the longitudinal nature of our dataset to study such changes. We only found a weak trend of increase of perceived expertise, possibly linked to the increased perception of technological quality in the six-year period captured in our study. On the other hand, no temporal patterns were found for the rest of the studied variables. Overall, the temporal dimension does not completely explain the variability of impressions. Future work needs to investigate additional sources of variability in more depth, and with a larger population of vloggers.

Finally, RQ3 asked about the possible benefits of inferring perceived traits and other attributes from multiple videos of a given user. Our results suggest that inferences at the user level, i.e., based on multiple videos from the same vlogger, achieve higher performance than inferences about a vlogger based on a single video. This result, while intuitive, adds to the current literature on recognition of perceived traits of social video users, which so far has addressed the problem based on single samples of behavior. This could also motivate the study of certain traits that might be inferred to some degree using a collection of videos for a given user, while remaining elusive to infer based on a single video. This is a possible topic for future research.

Acknowledgments

This work was partly supported by the NTT-Idiap Social Behavior Analysis Initiative (NISHA) and by the Swiss National

Science Foundation through the UBImpressed project, while the second and third authors were affiliated with Idiap. We thank Radu Negoescu and Joan-Isaac Biel (Idiap) for help with data collection and annotation.

REFERENCES

1. N. Ambady, F.J. Bernieri, and J.A. Richeson. 2000. Toward a Histology of Social Behavior: Judgmental Accuracy from Thin Slices of the Behavioral Stream. *Advances in Experimental social psychology* 32 (2000), 201–257.
2. Carl R Anderson. 1977. Locus of control, coping behaviors, and performance in a stress setting: a longitudinal study. *Journal of Applied psychology* 62, 4 (1977), 446.
3. Oya Aran and Daniel Gatica-Perez. 2013. One of a Kind: Inferring Personality Impressions in Meetings. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*. 11–18.
4. Anne Archambault and Jonathan Grudin. 2012. A longitudinal study of facebook, linkedin, & twitter use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2741–2750.
5. Joan-Isaac Biel, Oya Aran, and Daniel Gatica-Perez. 2011. You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube. In *Proceedings of AAAI International Conference on Weblogs and Social Media*.
6. Joan-Isaac Biel and Daniel Gatica-Perez. 2012. The Good, the Bad, and the Angry: Analyzing Crowdsourced Impressions of Vloggers. In *Proceedings of AAAI International Conference on Weblogs and Social Media*.
7. Joan-Isaac Biel and Daniel Gatica-Perez. 2013. The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs. *Multimedia, IEEE Transactions on* 15, 1 (Jan 2013), 41–55.
8. Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. FaceTube: Predicting Personality from Facial Expressions of Emotion in Online Conversational Video. In *Proceedings of ACM International Conference on Multimodal Interaction*. 53–56.
9. Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi YouTube!: Personality Impressions and Verbal Content in Social Video. In *Proceedings of ACM on International Conference on Multimodal Interaction*. 119–126.
10. Paul Boersma. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5, 9/10 (2002), 341–345.
11. Steven P Brown, William L Cron, and John W Slocum Jr. 1997. Effects of goal-directed emotions on salesperson volitions, behavior, and performance: A longitudinal study. *The Journal of Marketing* (1997), 39–50.
12. Laura E Buffardi and W Keith Campbell. 2008. Narcissism and social networking web sites. *Personality and social psychology bulletin* 34, 10 (2008), 1303–1314.
13. Jean Burgess and Joshua Green. 2009. *YouTube: Online Video and Participatory Culture*. Polity Press.
14. Robert A Emmons. 1987. Narcissism: theory and measurement. *Journal of personality and social psychology* 52, 1 (1987), 11.
15. Jennifer Flashman. 2012. Academic Achievement and Its Impact on Friend Dynamics. *Sociology of Education* 85, 1 (2012), 61–80.
16. Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in personality* 37, 6 (2003), 504–528.
17. Yağmur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven, and Rob van Lier. 2016. Deep Impression: Audiovisual Deep Residual Networks for Multimodal Apparent Personality Trait Recognition. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, 349–358.
18. Nicholas S Holtzman, Simine Vazire, and Matthias R Mehl. 2010. Sounds like a narcissist: Behavioral manifestations of narcissism in everyday life. *Journal of Research in Personality* 44, 4 (2010), 478–484.
19. Jina Huh, Leslie S. Liu, Tina Neogi, Kori Inkpen, and Wanda Pratt. 2014. Health Vlogs As Social Support for Chronic Illness Management. *ACM Trans. Comput.-Hum. Interact.* 21, 4, Article 23 (Aug. 2014), 31 pages.
20. Oliver P John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research* 2, 1999 (1999), 102–138.
21. Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, Raul Chandias Ferrari, Mehdi Mirza, David Warde-Farley, Aaron Courville, Pascal Vincent, Roland Memisevic, Christopher Pal, and Yoshua Bengio. 2016. EmoNets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* 10, 2 (01 Jun 2016), 99–111.
22. Patricia G Lange. 2014. *Kids on Youtube: Technical Identities and Digital Literacies*. Left Coast Press.
23. Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. 2012. Connecting Meeting Behavior with Extraversion: A Systematic Study. *IEEE Trans. Affect. Comput.* 3, 4 (Jan. 2012), 443–455.
24. Gil Levi and Tal Hassner. 2015. Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. 503–510.

25. Mattias R. Mehl and Tamlin S. Conner (Eds.). 2014. *Handbook of Research Methods for Studying Daily Life*. The Guilford Press.
26. Norman Miller, Geoffrey Maruyama, Rex J Beaber, and Keith Valone. 1976. Speed of speech and persuasion. *Journal of Personality and Social Psychology* 34, 4 (1976), 615.
27. Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011a. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In *Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI '11)*. 169–176.
28. Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011b. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*. ACM, 169–176.
29. Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep Multimodal Fusion for Persuasiveness Prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI 2016)*. 284–288.
30. Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational Analysis of Persuasiveness in Social Multimedia: A Novel Dataset and Multimodal Prediction Approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 50–57.
31. Delroy L Paulhus and Kevin M Williams. 2002. The dark triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of research in personality* 36, 6 (2002), 556–563.
32. Alex Pentland and others. 2008. Honest Signals: How They Shape Our World. *MIT Press Books* 1 (2008).
33. J Ashwin Rambaran, Andrea Hopmeyer, David Schwartz, Christian Steglich, Daryaneh Badaly, and René Veenstra. 2017. Academic functioning and peer influences: A short-term longitudinal study of network–behavior dynamics in middle adolescence. *Child development* 88, 2 (2017), 523–543.
34. Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203 – 212.
35. Jeffrey A Roberts, Il-Horn Hann, and Sandra A Slaughter. 2006. Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects. *Management science* 52, 7 (2006), 984–999.
36. Dairazalia Sanchez-Cortes, Oya Aran, Marianne Schmid Mast, and Daniel Gatica-Perez. 2012. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia* 14, 3 (2012), 816–832.
37. Dairazalia Sanchez-Cortes, Joan-Isaac Biel, Shiro Kumano, Junji Yamato, Kazuhiro Otsuka, and Daniel Gatica-Perez. 2013. Inferring mood in ubiquitous conversational video. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 22.
38. Holly B Shakya and Nicholas A Christakis. 2017. Association of Facebook use with compromised well-being: a longitudinal study. *American journal of epidemiology* 185, 3 (2017), 203–211.
39. Michelle N Shiota, Dacher Keltner, and Oliver P John. 2006. Positive emotion dispositions differentially associated with Big Five personality and attachment style. *The Journal of Positive Psychology* 1, 2 (2006), 61–71.
40. Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.
41. Michael Strangelove. 2010. *Watching YouTube: Extraordinary Videos by Ordinary People*. University of Toronto Press.
42. A. Subramaniam, V. Patel, A. Mishra, P. Balasubramanian, and A. Mittal. 2016. Bi-modal First Impressions Recognition using Temporally Ordered Deep Audio and Stochastic Visual Features. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing.
43. YouTube. 2018. YouTube Global Reach. `https://www.youtube.com/yt/about/press/`. (2018). Online; accessed May 2018.
44. Chen-Lin Zhang, Hao Zhang, Xiu-Shen Wei, and Jianxin Wu. 2016. Deep Bimodal Regression for Apparent Personality Analysis. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, 311–324.