The Good, the Bad, and the Angry: Analyzing Crowdsourced Impressions of Vloggers

Joan-Isaac Biel and **Daniel Gatica-Perez**
Idiap Research Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
{jibiel,gatica}@idiap.ch

Abstract

We address the study of interpersonal perception in social conversational video based on multifaceted impressions collected from short video-watching. First, we crowdsourced the annotation of personality, attractiveness, and mood impressions for a dataset of YouTube vloggers, generating a corpora that has potential to develop automatic techniques for vlogger characterization. Then, we provide an analysis of the crowdsourced annotations focusing on the level of agreement among annotators, as well as the interplay between different impressions. Overall, this work provides interesting new insights on vlogger impressions and the use of crowdsourcing to collect behavioral annotations from multimodal data.

Introduction

Given the evidence that suggests that social media content conveys information suitable to build personal impressions from users (Gosling, Gaddis, and Vazire 2007), research on interpersonal perception has investigated the formation of personality impressions based on user profiles (Gosling, Gaddis, and Vazire 2007) and blogs (Li and Chignell 2010). While these works provide understanding about the use of personal information, photos, and text as drivers for self-presentation and impression formation, relatively little is known about interpersonal perception in social video, in which vlogging is an increasingly important format. In addition, while these efforts have mainly investigated on personality impressions, a number of social psychology works have emphasized the interplay between personality judgments and other personal and social constructs such as physical attributes, intelligence or emotionality (Dion, Pak, and Dion 1990), that are also relevant variables in social media.

Some recent research has studied personality impressions in conversational vlogging on the basis of the nonverbal behavior displayed in the videos (Biel, Aran, and Gatica-Perez 2011). Conversational vlogging is a unique scenario for the study of impressions in social media as vloggers display in front of the camera rich, personal, spontaneous, audiovisual information that conveys both appearance and behavioral cues that may be useful to characterize vloggers independently of the verbal video content (i.e., what they say).

In this paper, we investigate the use of crowdsourcing to collect joint personality, attractiveness, and mood impressions for a dataset of conversational vlogs. Our study explores multiple types of impressions that can be formed on this social media setting, and addresses a broader number of traits and states than previous work (Biel, Aran, and Gatica-Perez 2011). In addition to its relevance from the interpersonal perspective, these type of crowdsourced corpora has also potential for the development of machine learning techniques to characterize vloggers, thus complementing work done with users from other social media spheres (Mishne 2005). Finally, our work contributes to efforts that explore the feasibility of using crowdsourcing to conduct human behavioral studies as well as a fast and affordable method of annotation that can scale to large amounts of social video and that collects impressions from demographically diverse annotators (Ross et al. 2010).

Our paper has two main contributions. First, we present an analysis of multifaceted crowdsourced impressions of vloggers with a central focus on the level of impression agreement achieved by MTurk workers. Second, we analyze the interplay between personality, attractiveness, and mood impressions, as well as how overall judgments of attractiveness and mood are made. Our paper shows that annotators achieve substantial agreement on their judgments and that several results from interpersonal perception and social psychology replicate on impressions made from crowdsourced, online video-watching.

Crowdsourcing Vlogger Impressions

We used Mechanical Turk to crowdsource the annotation of personality, attractiveness, and mood impressions from a dataset of conversational YouTube vlogs (Biel and Gatica-Perez 2011). To bound the cost of the collection, we limited the annotation task to a subset of one-minute conversational segments from 442 different vloggers. The exact process followed to obtain these segments is not explained here for space reasons but can be found in (Biel, Aran, and Gatica-Perez 2011). Our Human Intelligence Task (HIT) presented one single vlog segment on an embedded video player followed by three short questionnaires. With the purpose of obtaining spontaneous impressions, we did not give any par-

Questionnaire	Trait
Personality	Big-Five: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, Openness to Experience
Attractiveness	Beautiful, Likable, Friendly, Smart, Sexy, Overall attractiveness
Mood	Happy, Excited, Relaxed, Sad, Bored, Disappointed, Surprised, Nervous, Stressed, Angry, Overall mood
Demographics	Gender, Age (<12, 12-17, 18-24, 25-34, 35-50, >50), Ethnicity (Caucasian, Black/African American, Asian/Pacific Islander, American Indian/Alaskan native, Hispanic, Other)

Table 1: Summary of the crowdsourced annotations.

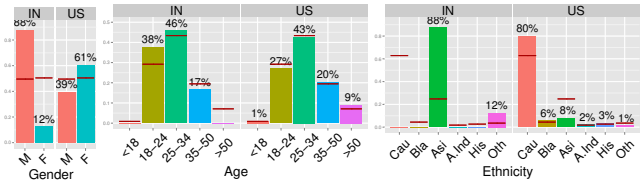

Figure 1: Demographic distribution (in %) of MTurk workers from US (N=89) and India (N=27). The superposed red lines indicate the distribution of the overall sample.

ticular instructions to workers on how to complete the HIT apart form 1) watching the video entirely and 2) answering the questions. We introduce the questionnaires as follows and summarize the annotated traits in Table 1.

Personality questionnaire We used the Ten-Item Personality Inventory (TIPI) (Gosling, Gaddis, and Vazire 2007), which was designed for settings with limitations on the time that people can spend answering questions. The form characterizes the Big Five personality traits on the basis of 10 items (two items per trait) and two adjectives per item. The rater is asked to judge the extent to which each pair of adjectives describes the target person on a 7-point likert scale.

Attractiveness questionnaire We built our own questionnaire inspired on research in attractiveness impressions (Fiore et al. 2008; Kniffin and Wilson 2004). First, we documented a list of attractiveness adjectives and then gathered them in pairs, as in the TIPI. The resulting questionnaire consisted of 5 items covering five different facets of physical and nonphysical attractiveness (one item per facet) and an additional 6th item to rate the overall attractiveness.

Mood questionnaire Based on a list of moods from the Livejournal blogging platform, used in other works (Mishne 2005), we built a list of twenty mood adjectives that we interpreted as possibly manifesting in vlogs. Then, we put together 10 items that cover ten different affective states (one item per state) of diverse arousal and valence, and we also added one last item to rate the overall mood.

Demographic questionnaire We asked MTurk workers to guess the gender, the age, and the ethnicity of vloggers, which is useful information to characterize users but is often missing from YouTube users' profiles. Because this annotation is clearly more objective that the requested in the questionnaires above, it also represents a useful proxy for measuring MTurk's work quality.

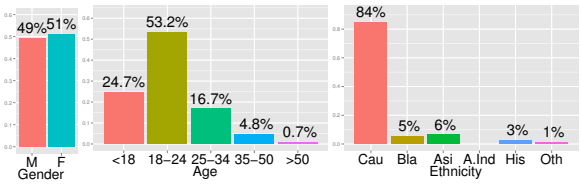

Figure 2: Vlogger demographics based on the majority voting answers from crowdsourced annotations.

In total, we posted 2,210 HITs in order to collect five different judgments for each of the 442 vloggers. We restricted the task to workers with HIT acceptance rates of 95% or higher, from the US (1,768 HITs) and India (442 HITs), as they are the English speaking countries with more workers (Ross et al. 2010). To sign up for the task, we asked annotators to self-report their demographics and to sign a consent of participation.

Analysis of Crowdsourced Impressions

Our analysis is structured in six sections. The first two sections provide some insights on the demographics of annotators and vloggers based on the crowdsourced data. In the third section, we focus on the evaluation of the agreement achieved among annotators on their impressions from vloggers. The level of agreement can be interpreted as a measure of annotation quality that helps to assess whether crowdsourcing is a suitable setting for annotating online social video. In addition, it is useful to identify what type of user traits are manifested and can be decoded in vlogging. In the fourth section, we investigate the correlations across impressions. In the fifth section, we study the formation of overall attractiveness and overall mood impressions. Finally, we report gender differences on building impressions.

MTurk Annotators Demographics

A total of 113 workers completed our HITs in MTurk. Figure 1 summarizes the demographics of MTurk workers, which illustrates the ease of obtaining a variate pool of annotators when using MTurk compared to gathering people offline. This diversity may be desirable to represent the variety of demographics found in online video audiences.

Our pool of annotators is balanced in gender but shows clear differences on the breakdowns between US (61% female and 39% male) and Indian workers (12% female and 87% male). Regarding age groups, we found most of MTurk workers on the ranges of 18-24 and 25-34, with indians younger than US workers. Finally, most of the US workers reported being Caucasian (80%), while most of indian workers reported themselves as Asian/Pacific Islander (88%). Interestingly, these demographic particularities resemble those reported by earlier investigations of the MTurk population demographics based on larger samples (Ross et al. 2010).

Crowdsourced Vlogger Demographics

Figure 2 shows the demographic distribution of YouTube vloggers based on the majority voting answer of MTurk workers. Our dataset of vloggers is mostly balanced in gender, and it is mainly constituted by vloggers below 24 years old and Caucasian. In addition to this basic insight, these demographic annotations provide a good opportunity to investigate the quality of annotations achievable in MTurk. We

Trait	$Mean$	SD	Min	Max	$Skew$	ICC
Extr	4.61	1.00	1.90	6.60	−0.32	.77
Agr	4.68	0.87	2.00	6.50	−0.72	.65
Cons	4.48	0.78	1.90	6.20	−0.32	.45
Emot	4.76	0.79	2.20	6.50	−0.57	.42
Open	4.66	0.71	2.40	6.30	−0.09	.47
Beautiful	4.41	1.02	1.40	6.80	−0.48	.69
Likable	4.98	0.80	2.20	7.00	−0.51	.44
Friendly	5.13	0.83	2.20	6.80	−0.67	.51
Smart	4.74	0.74	2.80	6.80	−0.19	.35
Sexy	4.06	1.14	1.00	7.00	−0.32	.60
Over. attract.	4.48	0.93	1.20	6.60	−0.49	.61
Happy	4.32	1.18	1.20	7.00	−0.39	.76
Excited	4.54	1.20	1.20	6.80	−0.39	.74
Relaxed	4.22	0.93	1.60	6.20	−0.50	.54
Sad	2.17	0.99	1.00	6.60	1.49	.58
Bored	2.41	1.04	1.00	6.80	1.20	.52
Disappointed	2.38	1.11	1.00	6.43	1.02	.61
Surprised	2.51	0.99	1.00	6.40	1.09	.48
Nervous	2.37	0.82	1.00	5.20	0.84	.25
Stressed	2.24	0.93	1.00	6.40	1.09	.50
Angry	2.15	1.10	1.00	6.60	1.68	.67
Over. mood	4.83	1.04	1.60	7.00	−0.58	.75

Table 2: Basic descriptive statistics of vlogger impressions: Mean, standard deviation (SD), minimum (min), maximum (max), skewness (Skew), and Intraclass Correlation Coefficients ICC(1,k). All ICCs are significant with $p < 10^{-3}$.

measured this quality by means of the Fleiss' Kappa coefficient, which assesses the reliability of categorical ratings and compensates for the agreement that could occur if raters were annotating at random. The Kappa coefficients resulted in high, fair, and moderate agreement for gender ($\kappa = .91$), age ($\kappa = .29$), and ethnicity ($\kappa = .46$), respectively, which concurs with the idea that the last two categories are more difficult to judge compared to gender. Clearly, the figures satisfy the level of agreement expected in a quality task.

Impressions' Statistics and Agreement

As a first step towards understanding the type of impressions collected from MTurk, we computed a set of descriptive statistics (Table 2). As observed from the minimum and maximum scores, all the annotations span fully across the 7-points likert scale, which indicates that all personality traits, attractiveness facets, and mood states are found in vlogs to some extent. The distribution of all personality traits, attractiveness facets, and of positive moods (Happiness, Excitement, and Relax) are centered on the positive side of the likert scales (≥ 4) and show little skewness ($\leq \pm 1$). In contrast, the rest of moods (negative and neutral) are centered low on the negative part of the scale and result positively skewed (≥ 1), which suggests that they may be displayed in conversational vlogs much less frequently than positive moods. As shown by the standard deviation, the dispersion of the scores also varied among annotations (typically larger variances of the aggregates are related to higher agreement between annotators).

We calculated the annotator agreement using Intraclass Correlation Coefficients ICC(1,k). Among personality traits, Extraversion and Agreeableness were the ones achieving the highest agreement. The first result may be not surprising, as Extraversion has been typically reported as the easiest trait to judge at first sight in many scenarios (Borkenau and Liebler 1992). However, compared to existing literature, finding Agreeableness as the trait with the second highest ICC seems particular to the vlogging scenario, and suggests that this setting may be more suitable than others to display or decode this trait. Regarding judgments of attractiveness, we found that the physical facets of attractiveness (Beautiful and Sexy) and the overall rating (Over. Attract.) reached levels of agreement comparable to Agreeableness. Several moods (e.g. Happiness, Excitement, Anger, Disappointment) as well as the Overall Mood achieve substantial annotator agreement. Interestingly, mood impressions were on average the judgements that achieved the highest agreement compared to personality and attractiveness. Overall, these ICCs provide valuable information in terms of the impressions that we can make from vlogging. However, it is unclear to what extent the low reliability of characteristics such as Smartness, Nervousness, or Surprise is due to the conversational vlogging setting itself, the duration of the vlog slices, or both.

	1	2	3	4	5	6	7	8	9	10	11	12
Extr												
Agr	.04											
Beautiful	.20	.30										
Friendly	.35	.57	.54									
Sexy	.17	.28	.82	.50								
Happy	.47	.38	.37	.52	.37							
Excited	.64	.26	.33	.49	.33	.74						
Relaxed	−.12	.40	.25	.37	.28	.34	.15					
Sad	−.39	−.32	−.15	−.34	−.12	−.36	−.37	−.10				
Bored	−.40	−.30	−.18	−.35	−.14	−.26	−.39	.03	.63			
Disapp	−.29	−.38	−.13	−.29	−.10	−.43	−.35	−.18	.74	.51		
Stressed	−.28	−.34	−.14	−.30	−.11	−.31	−.27	−.20	.71	.50	.68	
Angry	−.11	−.58	−.15	−.35	−.12	−.35	−.20	−.25	.56	.42	.67	.60

Table 3: Pair-wise correlations of selected judgements with ICC(1,k) > .50. (with the exception of values lower than $r = .10$, all correlations are significant with $p < 10^{-3}$).

Correlations between impressions

We evaluated the extent to which vlogger impressions were associated to each other by means of pair-wise correlations. For this analysis we focus on traits that showed substantial agreement (we choose those ICC(1,k) > .50 arbitrarily), and we did not include overall attractiveness and overall mood, which we address in the next subsection. Table 3 shows a number of significant correlations that may be explained by a well-documented halo effect that suggests that attractive people are typically judged as holding more positive traits than unattractive people, with some exceptions (Dion, Pak, and Dion 1990). For example, we found positive correlations between judgments of attractiveness and Extraversion (Beauty, $r = .20$, Friendliness, $r = .35$, and Sexiness, $r = .17$), which have been previously reported in the literature for other settings (Borkenau and Liebler 1992). In addition, we found that Beauty is positively correlated with positive moods (Happiness, $r = .37$, Excitement, $r = .33$, Relax, $r = .25$), and negatively correlated with negative moods (Sadness, $r = -.15$, Boredom, $r = . - 18$, Stress, $r = -.14$, and Anger, $r = -.15$). This halo effect may as well be mediating some of the correlations between Extraversion and moods (Happiness, $r = .47$ or Stress, $r = -.28$). It is important to note that compared to Extraversion, Agreeableness shows even stronger correlations with attractiveness

and mood (e.g. Beauty $r = .30$, Friendliness, $r = .57$, Happiness $r = .38$, Anger $r = -.58$), associations that may have not been observed in the literature because Agreeableness has achieved far less agreement in other scenarios (Borkenau and Liebler 1992). Also note that, while judgments of Extraversion and Agreeableness are not correlated, they show same sign effects with most of the attractiveness and mood scores with the exception of Relaxed, with whom they show opposite sign effects. We hypothesize that in the first case, Relaxed may have been interpreted as calmed (opposite to excited), whereas in the second case, it may have been judged as pleasant.

Overall Impressions of Attractiveness and Mood

We investigated the formation of overall attractiveness impressions on the basis of physical and nonphysical attractiveness by means of linear regression. We found that a combination of physical facets of attractiveness alone explained 77% of the overall attractiveness variance ($R^2 = .77$, $\beta_{beauty} = .50, t = 27.7, p < 10^{-3}$,$\beta_{Sexy} = .22, t = 14.8, p < 10^{-3}$), whereas a model of nonphysical facets explained 44% of the overall attractiveness variance ($R^2 = .44$, $\beta_{Likable} = .38, t = 13.1, p < 10^{-3}$, $\beta_{Friendly} = .16, t = 5.8, p < 10^{-3}$, and $\beta_{Smart} = .22, t = 10.3, p < 10^{-3}$). The use of stepwise linear regression procedures did not detect any optimal subset of judgments for none of the two models. We also tested the contribution of the nonphysical facets on judging overall attractiveness by comparing the physical attractiveness model to a full model that includes all facets (physical and nonphysical) using Analysis of Variance (ANOVA). The full model resulted to be significantly better than the physical attractiveness ($F = 10.2, p < 10^{-3}$) indicating that apart from physical facets, nonphysical facets contribute significantly to judgments of overall attractiveness, as it has been reported in the social psychology literature (Kniffin and Wilson 2004). The full model explained 80% of the attractiveness variance.

Regarding mood, we found that a linear regression explained 64% of the variance of the overall mood. The model, resulting from a stepwise procedure, included main contributions from Happiness ($\beta_{Happy} = .33, t = 18.7, p < 10^{-3}$), Excitement ($\beta_{Excited} = .20, t = 12.7, p < 10^{-3}$), Relax ($\beta_{Relaxed} = .19, t = 15.3, p < 10^{-3}$) and Anger ($\beta_{Angry} = -.16, t = -10.8, p < 10^{-3}$), and small yet significant contributions from Surprise ($\beta_{Surprised} = .08, t = 6.9, p < 10^{-3}$) and Stressed ($\beta_{Stressed} = -.03, t = -1.9, p < 10^{-2}$).

Gender Differences on Impressions

We explored whether impressions differed depending on the gender of annotators using one-way ANOVA tests. We summarize the significant effects as follows.

We found significant annotator gender effects for all personality trait judgments: mean personality scores given by female raters were higher than scores given by male raters (mean values are not reported for space reasons). Regarding personality and vlogger gender, we found significant effects for Agreeableness only: female vloggers scored higher on this trait than male vloggers.

We also found significant effects of both annotator gender and vlogger gender on judgments of attractiveness: female raters consistently gave higher ratings than males for all facets except for Sexual attractiveness, while female vloggers typically scored higher than male vloggers for all facets except Smart. In addition, by replicating the ANOVA experiments of the previous section for scores from same annotator gender, we also found that the contribution of nonphysical facets to a full linear regression model of overall attractiveness was larger for female annotators ($F = 43.8, p < 10^{-3}$) than for males ($F = 29.03, p < 10^{-3}$).

Finally, we found that male raters gave significantly higher scores than female raters for positive (except Excitement and Relax), as well as for negative moods and overall mood. In addition, we found vlogger gender effects for Happiness (female vloggers scored higher) and Anger (male vloggers scored higher).

Conclusions

We presented an original investigation on crowdsourced multifaceted human impressions on a dataset of conversational vlogs. Our work contributes new findings on interpersonal perception in social media and contributes to research exploring the suitability of crowdsourcing the annotation of multimedia corpora with human judgments. As a main relevant result, our study suggests that crowdsourcing is suitable for collecting interpersonal impressions from vloggers. We also found that several results from social psychology replicate on impressions made from online social video-watching Future work will investigate ways to identify joint patterns of these impressions, and to exploit these data to build predictive models of vlogger impressions.

Acknowledgements. This research was funded by the Swiss NSF through the NCCR IM2.

References

Biel, J.-I., and Gatica-Perez, G. 2011. Vlogsense: Conversational behavior and social attention in youtube. *TOMCCAP* 7(1):33:1–33:21.

Biel, J.-I.; Aran, O.; and Gatica-Perez, D. 2011. You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube. In *Proc. of ICWSM*.

Borkenau, P., and Liebler, A. 1992. Trait inferences: Sources of validity at zero acquaintance. *J . Per. Soc. Psych.* (62):645–657.

Dion, K. K.; Pak, A. W.; and Dion, K. L. 1990. Stereotyping physical attractiveness: A sociocultural perspective. *Journal of Cross-Cultural Psychology* 21(2):158–179.

Fiore, A.; Taylor, L.; Mendelsohn, G.; and Hearst, M. 2008. Assessing attractiveness in online dating profiles. In *Proc. of SIGCHI*.

Gosling, S. D.; Gaddis, S.; and Vazire, S. 2007. Personality impressions based on Facebook profiles. In *Proc. of ICWSM*.

Kniffin, K. M., and Wilson, D. S. 2004. The effect of nonphysical traits on the perception of physical attractiveness: Three naturalistic studies. *Evolution and Human Behavior* 25(2):88 – 101.

Li, J., and Chignell, M. 2010. Birds of a feather: How personality influences blog writing and reading. *International Journal of Human-Computer Studies* 68(9):589602.

Mishne, G. 2005. Experiments with mood classification in blog posts. In *Proc. SIGIR Workshop SATIA*, 19.

Ross, J.; Irani, L.; Silberman, M. S.; Zaldivar, A.; and Tomlinson, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proc. of CHI*.