Vlogcast Yourself: Nonverbal Behavior and Attention in Social Media

Joan-Isaac Biel
jibiel@idiap.ch

Daniel Gatica-Perez
gatica@idiap.ch

Idiap Research Institute
Ecole Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

ABSTRACT

We introduce vlogs as a type of rich human interaction which is multimodal in nature and suitable for new large-scale behavioral data analysis. Vlog analysis is useful not only to social media, but also to remote communication scenarios, and requires the integration of methods for multimodal processing and for social media understanding. Based on works from social psychology and computing, we first propose robust audio and visual cues to measure the nonverbal behavior of vloggers in their videos, and we then study the relation between behavior and the attention videos receive in YouTube. Our study shows significant correlations between some nonverbal behavioral cues and the average number of views per video.

1. INTRODUCTION

Conversational video blogs (vlogs) have evolved from a "chat from your bedroom" initial format to a highly creative form of expression and communication, resulting in a predominant type of user-generated video content on the Internet. While recent research in social media (including personal websites, blogs, and online social networks) has focused so far on the automatic analysis of text, and ethnographic studies have investigated some of the processes of creation and interaction through vlogging, we do not know of any previous attempts to analyze conversational vlogs automatically.

In this article, we introduce a new research domain in social interaction computing, namely the automatic analysis of human behavior in conversational vlogs. In short, the goal of this domain is the understanding of the processes involved on this hugely popular social media type, based not only on the patterns of contextual behavior of vloggers around their videos (e.g. uploads, views, social-oriented features), but on the specific ways vloggers behave in them. This research is not only relevant to understand this type of social media, but also contributes to the larger social interaction modeling agenda by studying a real-life communication scenario that is rich and complex, and that provides behavioral data at scales that have not been previously achievable due to natural limitations in other communication scenarios (e.g. in face-to-face dyadic or group interaction). Furthermore, compared the use of controlled face-to-face recordings [**?**, 5], vlogging analysis requires the integration of methods for both robust (yet simple) multimoal processing and for social media understanding.

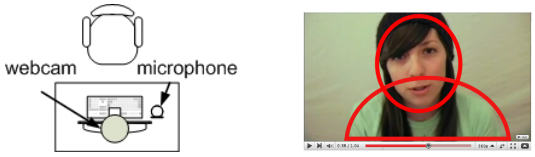

Figure 1: The basic formal rules of a vlog entry: a camera, a microphone (left), and a *talking head* (right).

Our article has mainly four contributions. First, we cast vlogging as a novel research domain as compared to other several studied types of real-life multimodal human interaction. Second, through the study of vlogging, we bring together nonverbal behavior analysis and social media analysis, going beyond the use of words and using an alternative communication channel. Third, we propose the use of robust audio and visual features to characterize vloggers that are motivated by social psychology, extracted automatically, and applicable at large scale. Finally, we present the first study of the relation between automatically extracted multimodal nonverbal behavior and social attention (measured by the number of views) in a sample of 2200 vlogs extracted from YouTube.

2. INTERACTION THROUGH VLOGS

In this work, we refer to vlogs as the original video-counterpart of text blogs that emerged with the advent of YouTube and other video-sharing sites, and that serve both as a life documentary and as a tool for communication and interaction on the Internet. Users of social media create, upload, watch, and share video content in a wide variety of formats that are loosely categorized as vlogs.

In their most basic format, vlogs are *conversational* videos, where people (as shown in Figure 1, usually a single person in the form of a *talking head*) discuss facing the camera and *addressing* the audience during most of the time in a Skype-style fashion. We are interested in this setting as it represents the simplest vlogging scenario and the one that features most conversational behavior (compared to other vlogging styles that might feature music playing, mashups, etc.). Furthermore, this type of vlog might be thought as the "direct" multimodal extension of traditional text-based blogging, where spoken works (i.e. what is said) are enriched

by the complex nonverbal behavior displayed in front of the camera. Conversational vlogs clearly share some features with other talking-head-type media such as professional or personal videoconferencing [13]. However, some fundamental differences are the asynchronous nature of vlogging and its "monologue" character. Moreover, the availability of a huge amount of metadata associated to the ' broadcast yourself reach everyone' core idea of YouTube, contrasts with private aspect of most video conferences.

3. YOUTUBE DATASET

We initially considered YouTube's existent content taxonomy[1] and we explored the distribution of vlogs among the different categories. We observed that vlogs were present in almost all the categories of users and videos, and that no specific category displayed a significant larger fraction of vlogs than the rest. Furthermore, we noticed different patterns on the frequency of vlog posts among users. For some, vlogging is a core activity that is practiced regularly (which results in a substantial collection of vlogs per user), whereas for others vlogging becomes a complement to other "main" activities, such as producing music videos.

For this study, we gathered a dataset of vlogs extracted from YouTube. Alternatively, we decided to query videos from YouTube using three possible keywords: "vlog", "vlogging", and "vlogger". In order to select conversational vlogs, we introduced a manual annotation task conceived to exploit the video collections of users. First, we extracted a list of 878 different *usernames* from video query results retrieved on November 17th 2009. Then, we recruited 10 untrained volunteers (whose only requirement was to be familiar with YouTube as a video viewer) that annotated (up to) the last 8 videos of each user, a total of 6396 videos. For the simply purpose the task, we explicitly recommended annotators to browse the videos using the progress bar, instead of watching them completely. Typically, each person spend one hour to annotate the videos corresponding to 25 vloggers.

Based on the annotations, we identified a final set of 2269 videos from 469 users. Our dataset contains all the videos together with their metadata (title, description, duration, keywords, video category, date of upload, number of views, comments, ratings, times "favorited", and average rating). Typical durations of vlogs are between 1 and 6 min (70% of the videos appear in this interval), with a median duration of 3.4min. Only 2.4% of the videos are longer than 10min, a limitation that can be only exceeded by certain users, called partners, which participate in the advertising scheme of YouTube. Once individual vlogs are aggregated over each user, this corresponds to more than 7min of video per vlogger for 80% of the vloggers in the collection (only 6% have less than 2min), which is a reasonable amount of "thin-slice" behavioral data. The concept of analyzing behavior based in brief observations ("thin-slices") has gained interest both in cognitive science [1] and social computing [15]. Overall, our dataset contains 151 hours of video.

4. AUTOMATIC PROCESSING OF VLOGS

Because of the unconstrained nature of vlogging, and the ease of using video editing software, vlogs result in extremely diverse content [12]. To the obvious diversity on audio volume, image quality, ligthing, etc. captured by the sensors, we must add vloggers practices' of including short video snippets (openings, closings, or sequences related to the conversation: outdoor scenes, pictures, recorded events, etc), which do not actually display the conversational setting described in Section 2. For the purpose of studying conversational interaction in vlogs, we discard these video snippets and then extract nonverbal behavioral cues on the conversational parts only. While complex techniques may exist for these purposes, the complexity of the content and their large-scale feasibility call for robust yet simple techniques. Figure 2 illustrates each one of the steps we followed to process vlogs.

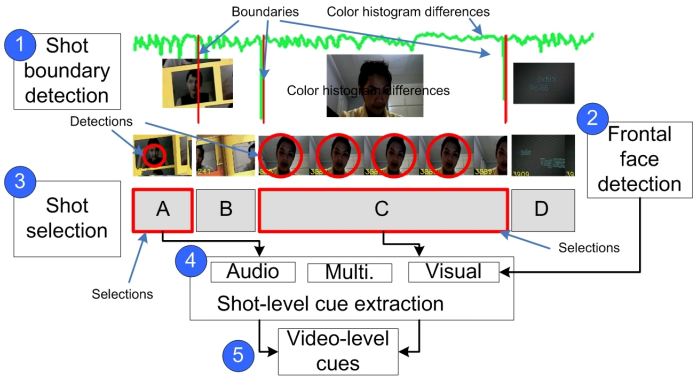

Figure 2: Automatic processing of vlogs.

4.1 Selecting conversational shots

We explored the use of several computer vision solutions for the purpose of detecting the conversational parts of vlogs, and among them, we choose a combination of a video shot boundary detector and a face detector (see ① and ② in Figure 2). First, we used the shot boundary detector to segment vlogs in different shots. Then, we selected shots (see ③ in Fig. 2) depending on the ratio of frames with face detections (to assess the presence of people) and the duration of the shot. The latter condition is motivated by the fact that video snippets interrumpt the main conversational scene tend to be short, independently on whether they feature people or not.

We used existing implementations based on the OpenCV library [3]. The shot boundary detection computes the Bhattacharyya distance between RGB color histograms of consecutive frames, and detects discontinuities based on a threshold. The face detector implements the boosted classifiers and Haar-like features from the Viola-Jones algorithm [**?**]. We set up and evaluated both systems in a small, random sample from our dataset, which was manually labeled for that purpose. Details are not disccussed here for space reasons.

4.2 Measuring nonverbal behavior

We investigate a number of automatic nonverbal behavioral cues extracted from both audio and video that have been shown to be effective to characterize some social constructs related to conversational interaction in both the social psychology literature [11] and more recently in social computing research [15, 8]. Vocalic cues and motion are correlated with levels of interest, extroversion, confidence, and openess and are good predictors of dominance [10], status [16], influence [7] and the interaction outcome [6]. While vlogs are not face-to-face conversations, it is clear that vloggers often behave as if they were having a conversation with

[1] As of May 2010, YouTube classifies users in 7 types (e.g. "Directors", "Comedians", "Gurus", etc) and videos in 15 categories (e.g. "People & Blogs", "Comedy", "Howto & Style", etc).

their audience. Thus, we hypothesize that the processes of nonverbal communication continue to exist in vlogging, and that they have a consequent effect on the interaction.

We extract the nonverbal cues for each selected shot and computed the average across shots to aggregate them into single video-level cues (see ④ and ⑤).

Audio cue extraction

We automatically extracted the cues using the toolbox developed by the Human Dynamics group at MIT Media Lab [14], which has proven to be robust to multiple conversational situations [15]. These cues are based on the voice/unvoiced and speech/non-speech segementations obtained from a two-level hidden Markov model (HMM) [2].

- **Speaking time.** Ratio between the total duration of the speech segments and the duration of the video.
- **Average length of speech segments.** Mean of the duration (in seconds) of all the speech segments. This measure relates to the frequency at which speech and pauses are produced (long segments equal to few pauses).
- **Voicing rate**. Ratio between the number of voicing segments and the total duration of the speech segments. It measures the speed at which a speaker articulates phonemes during a burst of speech (i.e how fast a persons speaks).
- **Speaking energy variation**. Energy's coefficient of variation in speech-only segments: the standard deviation divided by the mean. It is a measure of how well a speaker controls loudness.

Visual cue extraction

Some works have extracted visual activity cues by using features related to body motion, using visual focus of attention detection, or using hand or motion history techniques.

Instead, we explore the use of the face detector output (detection/non-detection, position and bounding box size) as a rough proxy for real gaze. For this purpose, we assume that frontal face detections occur when the vlogger looks towards the camera, as opposed to non-detections. Hence, we can obtain a looking/non-looking segmentation of the video frames to derive measures such as... We tried several measures, here we show some of them.

- **Distance to the camera.** Mean size of the bouding box across video frames normalized by the size of the frame (frames in a shot with no detections are discarded). Small ratios correspond to larger distances to the camera.
- **Looking time.** Ratio between the total duration of the looking segments and the duration of the video.
- **Looking rate.** Ratio between the number of looking segments and the total duration of the looking segments.
- **Head motion (1).** Standard deviation of the bouding box size (normalized by the frame size).
- **Head motion (2).** Coefficient of variation of euclidean distance between bounding box center and frame centers.

The first motion measure only captures planar movements only (vertical or horizontal), whereas the second also captures movements on the direction of the camera.

Video edition

Finally, we explore the effect of video edition itself, by considering features such as the relative number of shots, the video length, and the average length per shot.

5. NONVERBAL BEHAVIOR & ATTENTION

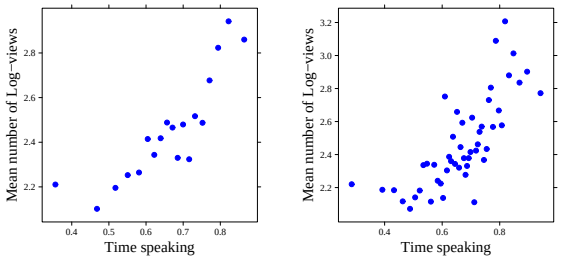

Figure 3: Median number of log-views received by vlogs with different speaking times. (we use 20 and 50 bins for the left and right plots, respectively).

Attention vs. popularity

Social media analysis have typically taken the number of views as a reference measure of video popularity, because it reflects the number of times that video has been accessed, resembling the way audiences are measured in traditional mainstream media [4]. Then, taking speaking time as a case of example, one may hypothesize that a vlogger talking more that an other, will recive proportionally more views (cite our second paper). In this paper, we define the average level of attention of a set of videos, as the average of all their views. We see these two measures as having different granularity. Whereas popularity accounts for a fine-grained measure on the number of views of videos (which follow a power-law distribution and is result of a combination of several complex mechanisms, such as preferential attachment and information filtering []), the level of attention is a more coarse measure that may be useful to explain broader patterns.

Analysis of correlation

We used standard Pearson's correlation measure between nonverbal behavioral cues and the average number views. For this purpose, we first group videos in equally-filled bins depending on the measure of the nonverbal cue (e.g. speaking time), and then compute the average number of views for each bin. This methodology is inspired by a procedure used in a recent study on the effect of attention on the patterns of content production in social media [9].

As an example, Figure 3 shows the average number of views received by vlogs with different speaking times. The correlation between both variables is 0.67 ($p < 10e\text{-}7$, bins=50). One could argue that the correlations computed are valid if the distributions of views for the bins are significantly different. To test this condition, we conducted of a Welch's test of the null hypothesis H1: "The distributions of the bins are the same". Welch's test is an adaptation of Student's t-test which does not assume the variances to be equal. We performed the test for numbers of bins between 10 and 100 and obtained p-values between lower than 0.001, which suggests that the hypothesis can be rejected.

6. CORRELATION RESULTS

Table **??** shows the correlation values for the different nonverbal cues from and the average number of log-views per video.

Audio modality

The correlation tests indicate that the speaking time, the average length of speech segments and the voicing rate are positively correlated with attention. This is, vloggers talking more, faster, and using few pauses receive, in mean, more views. On the other hand, the variation on speaking energy is negatively correlated to attention, which indicates that indeed, vocal control has also an effect on the way vlog-

Feature	*Corr*
Audio cues	
Speaking time	$.75^{***}$
Avg. length of speech segments	$.71^{***}$
Voicing rate	$.30^{*}$
Variation of speaking energy	$-.32^{*}$
Video cues	
Distance to the camera	$-.59^{**}$
Looking time	$.59^{***}$
Looking rate	$-.47^{***}$
Head motion (1)	$-.41^{*}$
Head motion (2)	$-.63^{***}$
Video edition	
shot ratio	$.29^{*}$
length	.15

Table 1: Pearson's correlation between visual cues and average number of views. $^{*}p < .01$,$^{**}p < .001$,$^{***}p < .0001$

gers are perceived in social media. Interestingly, our results compare to findings on face-to-face interactions, where for example, these cues were predictors of success on salary negotiations [6].

Visual modality

We also obtained strong correlations for several visual cues. Our analysis suggests that respecting an "optimal" distance with respect to the camera may have an effect on the communication process in vlogging, which penalizes those being too close to the camera. Whreas the looking time shows a positive correlation with the level of attention of vlog posts (as with speaking time) the frequency at which the vlogger interrupts his visual contact (the looking rate) has a negative impact on attention. Finally, both measures of motion revealed a negative correlation with attention. This is an interesting result, because other works have suggested that successful people in meeting interactions tend to be more active []. We hypothesize that these two features may capture specific patterns of head movement, and that other measures of motion (e.g. based on gesture) could show different results.

Video edition

Whereas the number of shots (taken somehow as measure of the level of video edition used) shows low correlation with attention. the lenght of the video does not show any significant correlation.

7. DISCUSSION

Issues to discuss:

- What is the impact of shot selection (vs not shot selection)?
- What is the impact of the dynamics of views on the study?
- How to verify that isual cues are actually measuring what we say?
- Usefulness of findings?
- What is the effect of the number of bins on the analysis?
- What about addressing aspects of verbal behavior?

8. CONCLUSIONS

We introduced a new domain on social interaction computing, namely the automatic analysis of conversational vlogs, which is multimodal in nature and has potential for large-scale analysis. We presented a first, original study on the use of robust audio and visual techniques to select conversational parts and extract nonverbal behavior from vlogs. Our analysis in a sample of vlogs from YouTube, shows evidence that cues extracted from the videos such as time speaking, the time looking, and the distance with respect to the camera, are correlated with attention, and may play role in the communication process of vlogging.

Overall, our work on studying attention shows promising results for the automatic modeling of behavior from online video, and may lead to the study of other constructs such as personality, persuasion, etc. Further research may also be dedicated to study larger samples of data, modeling users instead of videos.

Acknowledgments: We thank the support of the Swiss National Center of Competence (NCCR) on Interactive Multimodal Information Management (IM)2 and the voluntary annotators.

9. REFERENCES

[1] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin*, 111(2):256–274, 1992.

[2] S. Basu. *Conversational scene analysis*. PhD thesis, September 2002.

[3] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, Cambridge, MA, 2008.

[4] J. Burgess and J. Green. *YouTube: Online video and participatory culture*. Polity, Cambridge, UK, 2009.

[5] T. Choudhury and A. Pentland. Sensing and modeling human networks using the sociometer, 2003.

[6] J. R. Curhan and A. Pentland. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3), 05 2007.

[7] N. E. Dunbar and J. K. Burgoon. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005.

[8] D. Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[9] B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *J. Inf. Sci.*, 35(6):758–765, 2009.

[10] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *Trans. Audio, Speech and Lang. Proc.*, 17(3):501–513, 2009.

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

[12] B. M. Landry and M. Guzdial. Art or circus? characterizing user-created video on youtube. Technical report, SIC Technical Reports, Georgia Institute of Technology, 2008.

[13] B. O'Conaill, S. Whittaker, and S. Wilbur. Conversations over video conferences: an evaluation of the spoken aspects of video-mediated communication. *Hum.-Comput. Interact.*, 8(4):389–428, 1993.

[14] A. Pentland. Social dynamics: Signals and behavior. In *in International Conference on Developmental Learning*, October 2004.

[15] A. Pentland. *Honest signals: How they shape our world*, volume 1 of *MIT Press Books*. The MIT Press, 2008.

[16] C. L. Ridgeway. Nonverbal behavior, dominance, and the basis of status in task groups. *Journal of Social and Personal Relationships*, 52(5):683–694, 1987.