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. The automatic analysis of vlogs is useful not only to study social media, but also 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, visual, and multimodal cues to measure the nonverbal behavior of vloggers in their videos. Then, we investigate 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. Recent research in social media (including personal websites, blogs, and online social networks) has focused so far on the automatic analysis of text [8]. In addition, ethnographic studies have investigated some of the processes of creation and interaction through vlogging [12]. However, 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 [7, 14], vlogging analysis requires the integration of methods for both robust yet simple multimodal processing and for social media understanding.

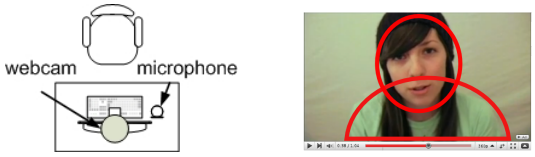

Figure 1: The basic formal rules of a vlog entry: a camera, a microphone (left), and a *talking head* (right).

Our article has mainly four contributions. First, we cast vlogging as a novel research domain, 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 analysis of words and studying an alternative communication channel. Third, we propose the use of robust audio, visual, and multimodal 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 media attention in a sample of 2200 vlogs extracted from YouTube.

2. INTERACTION THROUGH VLOGS

In this work, we refer to vlogs as the original audiovisual 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.

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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ICMI-MLMI'10 November 8-12, 2010, Beijing, China.
Copyright 2010 ACM 978-1-4503-0414-6/10/11 ...$10.00.

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

For this study, we gathered a dataset of vlogs extracted from YouTube. With the purpose of selecting vlogs featuring the conversational setting described above, we first queried videos from YouTube using three possible keywords: "vlog", "vlogging", and "vlogger", and then we introduced a manual annotation. First, we extracted a list of 878 different *usernames* from the video query results retrieved on November 17th 2009. Second, we recruited 10 untrained volunteers (whose only requirement was to be familiar with YouTube as a video viewer) who annotated up to the most 8 recent videos of each user, resulting in a total of 6396 videos. For this 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 finally identified a set of *mainly conversational* vlogs.

Our dataset contains 2269 videos from 469 users with their metadata (title, description, duration, keywords, video category, date of upload, number of views, and comments). Typical durations of vlogs are between 1 and 6min (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 for user, this corresponds to over 7min of video per vlogger for 80% of the vloggers in the collection, which can be seen as 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 [14]. Overall, our dataset comprises 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 [5]. To the obvious diversity on audio volume, image quality, lighting, 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. With the aim 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 this purpose, 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.

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

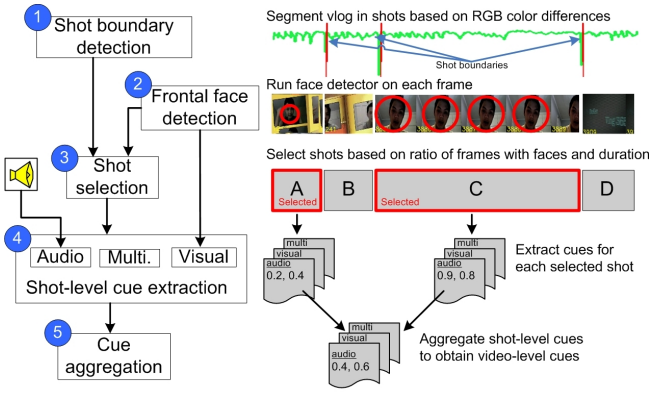

Figure 2: Automatic processing of vlogs.

boundary detector and a face detector (steps ① and ② in Figure 2). First, we used the shot boundary detector to segment vlogs in different shots. Then, we selected shots (step ③ in Figure 2) depending on the ratio of frames with face detections (to assess the presence of people) and the shot duration. The latter condition is motivated by the fact that video snippets interrupting the main conversational scene tend to be short, independently on whether they feature people or not.

We used existing implementation of both algorithms on the OpenCV library [4]. 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 [16]. We set up and evaluated both systems in a random sample of 100 vlogs from our dataset, which was manually labeled with that intention. Details are not disccussed here for space sake.

4.2 Measuring nonverbal behavior

We investigated a number of automatic nonverbal behavioral cues extracted from both audio and video that have been shown to be effective to characterize social constructs related to conversational interaction in both social psychology [11] and more recently in social computing [7, 14]. Vocalic and motion cues are correlated with levels of interest, extroversion, and openness, and are good predictors of dominance [10], status [15], and the interaction outcome [6]. In addition, we explored joint (multimodal) cues, which have also been studied in multi-party conversations [10]. While vlogs are not face-to-face conversations, it is clear that vloggers often behave as if they were having a conversation with 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 extracted a number of nonverbal cues for each selected shot as described in Section 4.1. and computed the average over shots to aggregate them into single video-level cues (steps ④ and ⑤ in Figure 2).

Audio cue extraction

We automatically extracted the audio cues using the toolbox developed by the Human Dynamics group at MIT Media Lab, which has proven to be robust to multiple conversational situations [14]. These cues are based on voiced/unvoiced and speech/non-speech segmentations 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 (longer segments relate to fewer 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 variation in speech-only segments, given by the standard deviation divided by the mean. It is a measure of how well a speaker controls loudness.

Visual cue extraction

Few works in conversational modeling have extracted visual activity cues related to hand or body motion, and to the visual focus of attention [10]. Alternatively, here we explore the use of the face detector output (detection/non-detection, face position and size) as a rough proxy for actual motion and gaze. For this purpose, we assume that frontal face detections occur when the vlogger looks towards the camera, which is reasonable in the typical vlog setting. As a result, we can obtain a looking/non-looking segmentation of the video frames which we use to derive some basic visual cues. Among the cues we explored, here we discuss some of them.

- **Distance to the camera.** Mean size of the face bounding box over video frames normalized by the frame size. 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 bounding box size normalized by the frame size.
- **Head motion (2).** Variation of the euclidean distance between bounding box center and frame center, given by the standard deviation divided by the mean. .

Whereas the first motion measure captures translational motion only (vertical or horizontal), the second also captures motion on the direction of the camera.

Multimodal cues

We explored a combination of multimodal features based on the speaking and looking segmentations. Using both segmentations we identified segments corresponding to 'looking while speaking' (L&S), 'looking while not speaking'(L&NS), and 'not looking while speaking' (NL&S). Then, we computed the **L&S/L&NS** and the **L&S/NL&S** ratios.

Video edition

Finally, as complementary features extracted from the video, we considered features such as the relative **number of shots** and the **video duration**. These can be seen as rough measures of video edition.

5. NONVERBAL BEHAVIOR & ATTENTION

Attention vs. popularity

Social media studies on video-sharing have typically considered the number of views received by a video as a reference measure of its *popularity*, because it reflects the number of times that specific item has been accessed, resembling the way audiences are measured in traditional mainstream media [5]. Alternatively, in this paper, we define the *average level of attention* of a set of videos as the average number of their views. Consider speaking time for example. One might hypothesize whether a vlogger talking more that another recives *proportionally* more views [3] (using the definition of popularity), or instead; whether vloggers talking more receive, in *average*, more views (using the definition of attention). We see these two measures as having different granularity. Whereas popularity accounts for a fine-grained measure on the exact number of views of videos, the level of attention is a more coarser measure that may be useful to explain broader patterns.

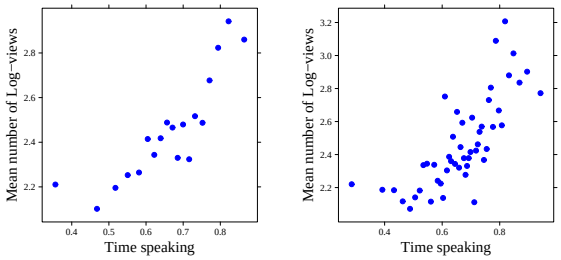

Figure 3: Median number of log-views received by vlogs with different speaking times (using 20 and 50 bins, from left to right).

Correlation analysis

We used the standard Pearson's correlation measure between nonverbal behavioral cues and the average number of 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 log-views for each bin. This methodology is inspired by a procedure used in a recent study of the effect of attention on the patterns of content production in social media [9].

As an example, Figure 3 shows the average number of log-views received by vlogs with different speaking times. The correlation between both variables is 0.75 ($p < 10^{-6}$, 50 bins). 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 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 lower than 0.001 in all cases, which suggests that the hypothesis can be rejected. We repeated this methodology for all nonverbal cues and we present results in the following section using 50 bins.

6. RESULTS

Table 1 shows the correlation values for the different nonverbal cues 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 average, more views. On the other hand, the variation on speaking energy is negatively correlated to attention, which suggests that, vocal control has also a relation with the way vloggers are perceived in YouTube. Interestingly, our results compare to findings on face-to-face interactions, where for example, these cues were predictors of success on salary negotiations [6].

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	$.29^{*}$
Duration	.15
Multimodal cues	
L&S/NL&S ratio	$.66^{***}$
L&S/L&NS ratio	$.32^{*}$

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

Visual modality

We also obtained significant correlations for several visual cues. The analysis suggests that following 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. Whereas 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 correlation with 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 visually active [7]. 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.

Multimodal cues

The two proposed measures, the 'looking while speaking'-'not looking while speaking' ratio (L&S/NL&S) and the 'looking while speaking'-'looking while not speaking' ratio (L&S/L&NS) show positive correlations with the average number of views. Multimodal cues based on speaking and looking turns have also been found to be effective in predicting dominance in multi-party conversations [10].

Video edition

Finally, the number of shots and the video duration show low and no significant correlation respectively. Interestingly, video duration has been typically a concern among vloggers [17].

7. DISCUSSION AND 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 study of the use of 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, the distance to the camera, and multimodal cues are correlated with the average number of views. Note that we do not claim any causality effect between these cues and social attention. We believe that our results provide initial evidence that, in addition to the content, nonverbal behavior plays a role in the communication process of vlogging and may affect how vloggers are perceived. Most likely, these nonverbal cues are related to social constructs such specific personality traits (like extroversion) or persuasion, and to how effective people are at creating vlogs. Thus, we aim to address these questions in future work.

In a more technical aspect, we acknowledge the need of validating that the proposed visual cues are indeed capturing those aspects of visual activity intended, and that are sufficient for that intent. Furthermore, we would like to study a larger sample of data, which will help to back up the significance of our findings. Finally, we might start addressing aspects of verbal behavior in vlogging, which to our knowledge have not either been studied.

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

8. REFERENCES

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

[2] S. Basu. *Conversational scene analysis*. PhD thesis, MIT Media Lab, Sept. 2002.

[3] J.-I. Biel and D. Gatica-Perez. Wearing a YouTube hat: directors, comedians, gurus, and user aggregated behavior. In *Proc. of ACM MM'09*, 2009.

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

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

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

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

[8] J. A. Gill, S. Nowson, and J. Oberlander. What are they blogging about? Personality, topic, and motivation in blogs. In *Proc. of AAAI ICWSM*, 2009.

[9] B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *Journal of Information Science*, 35(6), 2009.

[10] D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions Audio, Speech and Language Processing*, 17(3), 2009.

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

[12] H. Molyneaux, S. O'Donnell, K. Gibson, and J. Singer. Exploring the gender divide on YouTube: An analysis of the creation and reception of vlogs. *Journal of American Communication*, 10(2), 2008.

[13] B. O'Conaill, S. Whittaker, and S. Wilbur. Conversations over video conferences: an evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction*, 8(4), 1993.

[14] A. Pentland. *Honest signals: How they shape our world*. The MIT Press, 2008.

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

[16] P. Viola and M. Jones. Robust real-time object detection. *Int. Journal of Computer Vision*, 57(2), 2002.

[17] YouTube Blog. Upload limit increases to 15 minutes for all users, July 2010.