VlogSense: Conversational Behavior and Social Attention in YouTube

JOAN-ISAAC BIEL and DANIEL GATICA-PEREZ,
Idiap Research Institute and Ecole Polytechnique Fédérale de Lausanne (EPFL)

We introduce the automatic analysis of conversational vlogs (VlogSense, for short) as a new research domain in social media. Conversational vlogs are inherently multimodal, depict natural behavior, and are suitable for large-scale analysis. Given their diversity in terms of content, VlogSense requires the integration of robust methods for multimodal analysis and for social media understanding. We present an original study on the automatic characterization of vloggers' audiovisual nonverbal behavior, grounded in work from social psychology and behavioral computing. Our study on 2,269 vlogs from YouTube shows that several nonverbal cues are significantly correlated with the social attention received by videos.

Categories and Subject Descriptors: H.1.2 [**User/Machine Systems: Human Information Processing**] Information Processing; H.3.1 [**Information Storage and Retrieval**] Content Analysis and Indexing

General Terms: Human Factors, Measurement

Additional Key Words and Phrases: vlogging, YouTube, social media, nonverbal behavior

ACM Reference Format:

Biel, J-I. and Gatica-Perez, D. VlogSense: Conversational Behavior and Social Attention in YouTube. ACM Trans. Multimedia Comput. Commun. Appl. 2, 3, Article 1 (May 2010), 20 pages.
DOI = 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

1. INTRODUCTION

We are rapidly adopting social media as a natural form of interaction, enlarging the list of channels that we frequently use for communication in our daily life, just like we once did with e-mail and mobile phones. The popularity of blogging and social networking sites such as Wordpress, Twitter, or Facebook makes evident that text is still the most predominant form of online interaction. However, early adopters of a richer medium such as online video, commonly known as video bloggers (or in short, vloggers), have been experimenting with video as a form of effective communication for a few years. Conversational vlogs have evolved from a "chat from your bedroom" initial format to a highly creative form of expression and communication, resulting in a prevalent type of user-generated video content on the Internet [Burgess and Green 2009]. Today, video-sharing sites such as YouTube are not simply seen as a place for watching and sharing videos, but as a suitable platform for daily interaction, e-

This work is supported by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2.
Author's address: J-I. Biel, Idiap Research Institute, Centre du Parc, Rue Marconi 19, CP 592. CH-1920 Martigny, Switzerland.

DOI 10.1145/0000000.0000000 http://doi.acm.org/10.1145/0000000.0000000

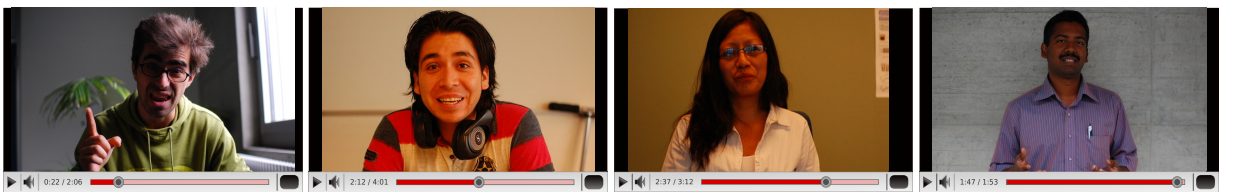

Fig. 1: Examples of conversational vlogs.

learning, product marketing, corporate communications, and other applications where a more natural and engaging way of reaching the audience is either necessary or might be beneficial [Burgess and Green 2009; Strangelove 2010; Vonderau 2010].

Recent research in social media has focused on the automatic analysis of text in personal websites, blogs, and social networks, primarily because of their popularity, but also because, compared to video, text is easier to process at large-scale and there are plenty of analytical tools available [Kramer and Rodden 2008]. These works have shown how specific patterns of text editing and expression, together with contextual metadata, are useful to understand users' motivations [Goswami et al. 2009; Kramer and Rodden 2008; Gill et al. 2009], emotions [Mishne 2005], and personality traits [Gill et al. 2009], as well as how users interact with others or how they are perceived by them [Gill et al. 2009]. Yet, there is a relatively little understanding about what aspects are important for an effective use of video as a social communication medium. Although a few ethnographic studies have investigated some of the underlying motivations, and the processes of creation and interaction through vlogging [Lange 2007], we do not know of any attempt to automatically analyze conversational vlogs using multimodal techniques that can extract actual displayed behavior and that are applicable at large-scale.

In this article, we introduce a new research domain in social media, namely the automatic analysis of human behavior in conversational vlogs (see examples in Figure 1 [1]). In short, the goal of this domain is the understanding of the processes involved in this hugely popular social media type, based not only on the verbal channel – what is said –, but also on the nonverbal channel – how it is said. The nonverbal channel includes prosody, gaze, facial expression, posture, gesture, etc. and has been studied in depth in the field of nonverbal communication [Knapp and Hall 2005]. This research is not only relevant to understand this specific type of social media, but also contributes to the larger social interaction modeling agenda [Gatica-Perez 2009; Pentland 2008], by studying a real-life communication scenario that is rich and diverse, and that provides behavioral data at scales that have not been previously achievable in other controlled communication scenarios (face-to-face dyadic conversations and group meetings). Furthermore, compared to the analysis of the above scenarios, vlogging analysis requires the integration of methods for social media understanding as well as for robust and tractable multimedia processing.

Our work has four contributions. First, we cast vlogging as a novel research domain in social media, compared to both text-based social media analysis and face-to-face interaction analysis. Second, through the automated study of vlogging, we bring together nonverbal behavior analysis and social media analysis, going beyond the analysis of words and examining an unexplored communication channel. The joint study of nonverbal behavior and social media has significant potential beyond vlogging, as new forms of rich, online communication channels (e.g. video calls, video chats) continue to emerge. Third, with the aim of characterizing vlogger behavior, we propose the use of robust, automatically extracted audio, visual, and multimodal nonverbal cues that are motivated by social psychology, and

[1] All the images used in this paper to illustrate vlog content show people that volunteered to pose as vloggers. In addition, the video playing bar was added as a mockup.

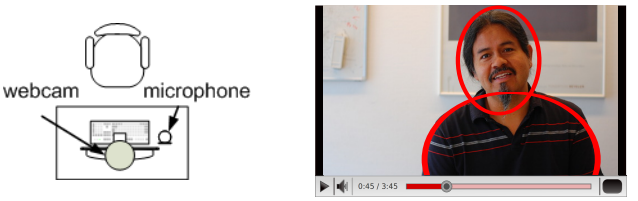

Fig. 2. The basic vlog setup: a camera, a microphone (left), and a talking-head (right).

applicable at large-scale. Finally, using a sample of over 2,200 vlogs extracted from YouTube, we provide novel insights on common patterns of vlogger behavior based on automatically extracted cues. In particular, we first hypothesize, and then show, that specific patterns of vloggers' behavior are correlated with the level of attention that vlogs receive from the YouTube audience. Social attention, which we measured on the basis of the video views, is a valuable resource in social media, as it is commonly seen as a public reward to users than contribute with quality content [Huberman et al. 2009].

We organized the rest of the article as follows. In Section 2, we describe the basic formal rules of vlogging. In Section 3, we review related work from different areas of social media and behavioral computing. In Section 4, we present our data collection. In Section 5, we introduce our approach to the automatic processing of vlogs. In Section 6, we present our experiments and discuss our findings. Finally, in Section 7, we summarize our research and provide future work directions.

2. VLOGGING: MULTIMODAL INTERACTION IN SOCIAL MEDIA

In this work, we focus on the study of *conversational vlogs*. Since vlogging is an umbrella term that may include a diversity of formats, it is important to understand the characteristics of conversational vlogging and our research interest in studying them, compared to other vlogging formats such as sketch-comedies, musical performances, or home scene footage.

In their most basic format, vlogs are conversational videos, where people, usually a single person in the form of a talking-head, discuss facing the camera and addressing the audience in a Skype-style fashion (see Figure 2). We are interested in this configuration for two reasons. First, compared to other types of vlogging, the single talking-head format is the one whose content is mostly conversational. the most conversational behavior. This is relevant to social media research, as this type of vlogs might be thought as the "direct" multimodal extension of traditional text-based blogs, where spoken words – what is said – are enriched by the complex nonverbal behavior displayed in front of the camera – how it is said. Second, there is evidence that conversational vlogs are a very popular format among user-generated video. In a recent study, Burgess and Green [2009] found the single talking-head format to account for 40% of the most popular content among the videos manually classified as "user-generated" on a sample of 4,000 videos extracted from YouTube. The popularity of this type of vlog evidences the availability of large-scale data.

Conversational vlogs also share some characteristics with other talking-head-type media such as professional videoconferencing and personal video calls. However, three fundamental differences are: the asynchronous nature of vlogs, their monologue-like nature, and the availability of a huge amount of data and metadata, which are associated with the "broadcast yourself, reach everyone" core idea of YouTube, and that clearly contrasts with the private aspect of most video conferences.

Although conversational vlogging is obviously not exclusive of YouTube, the forms of social engagement inherent in vlogging are key features that distinguish YouTube as a platform for creativity and participation around video, rather than just a video repository and distribution system [Burgess and Green 2009; Strangelove 2010]. Today, vlogs constitute a communication genre that promotes high participation, critique, and discussion, and therefore are not only used for life documentary or daily interaction, but also for e-learning, entertainment, marketing, and corporate communication, where a

more social (even personal) way of reaching the audience is either necessary or beneficial [Burgess and Green 2009; Strangelove 2010; Vonderau 2010].

Like in other social media, vloggers learn new competencies and innovate as they participate and interact in YouTube, which results in a wide diversity in the technical quality of vlogs. Quality depends not only on the equipment used, which is accessible and cheap today, but also on the extent to which users possess or develop the necessary skills to create video. For example, some vloggers may lack the skills needed to control technical aspects such as the intelligibility of the audio, or the design of the scene, or might simply ignore them. Furthermore, while some vloggers post one-take, raw scenes in front of the webcam, other vloggers upload edited video, consciously selecting excerpts of conversational footage, and add soundtracks, openings, endings, and other video snippets that are not necessarily conversational but that accompany, illustrate, or color their monologues.

3. RELATED WORK

Our study leverages research from both social media and behavioral computing. On one hand, our work represents a first attempt to study vlogger nonverbal behavior which complements methods in social media literature addressing the analysis of verbal content in traditional blogs and social networks [Kramer and Rodden 2008; Goswami et al. 2009]. On the other hand, our work extends research in behavioral computing by investigating ways to characterize and understand human nonverbal behavior in an unexplored type of interaction at large-scale.

YouTube has become a subject of research in a range of domains. A number of research works in computer science have treated YouTube as a video repository and distribution system, and studied the impact of user generated content in video-on-demand architectures [Cha et al. 2007; Cheng et al. 2008]. Research has also addressed YouTube as a platform for participatory culture, studying common patterns related to video production, interaction, and social participation. Focusing on *videos* as the unit of analysis, research has examined the type, the properties, and the editing constituents of YouTube videos, in order to gain understanding about user-generated media production [Landry and Guzdial 2008], and to investigate the role and impact of this content in YouTube as compared to traditional media content [Halvey and Keane 2007; Burgess and Green 2009]. Whereas these works relied on the manual coding of small datasets, few studies attempted to treat YouTube's content automatically, for example, to asses the topic and ideological perspective of videos based on their tags [Lin and Hauptmann 2008]. Focusing on *users* as the subject of analysis, research has automatically analyzed the metadata traces left by users in YouTube as a mean to understand the long-term behavior [Kruitbosch and Nack 2008; Biel and Gatica-Perez 2009] and production patterns of users [Huberman et al. 2009], and has also investigated their social interactions based on the friendship and subscriptions links between them [Mislove et al. 2007]. Our research is also focused on users, but differs from these works in that we study YouTube user behavior as it is actually displayed on the videos themselves, going beyond the use of video and user metadata. In addition, our work differs from other works that have used the term *videoblog* in the sense of "a log of videos", instead of a conversational vlog [Zhang et al. 2009].

Despite the high popularity of vlogging, we know of very few attempts that investigate the characteristics of vlogs as a rich communication medium, mainly through ethnographic studies. Though some of these research works conceptualized vlogs in a broader sense than the pure conversational one, in practice they consistently reported the major part of the content as featuring a single participant talking to the camera. In particular, works have focused on the specific problems of self-presentation [Griffith 2007], gender differences in the creation and reception of vlogs [Molyneaux et al. 2008], and how specific age groups interact and participate in YouTube [Lange 2007; Harley and Fitzpatrick 2009].

Overall, these works relied on qualitative observations based on verbal content, video comments, and personal interviews, and were therefore limited to small datasets. Our work differs in that we propose and demonstrate the feasibility of analyzing conversational vlogs, automatically processing and exploiting the large amount of data available online. In addition, we aim to study the nonverbal behavior of vloggers, as opposed to their verbal discourse, using objective, measurable features.

The automatic analysis of vlogs, as we present it here, extends within the multimodal domain the growing body of studies on social media focused on blogs and personal websites that seek to provide insights on user behavior. Research has analyzed text (i.e. word usage, linguistic styles) to study blogs' topics and bloggers' intentions [Kramer and Rodden 2008], to classify their mood [Mishne 2005], or to infer personality types [Gill et al. 2009] and demographics [Goswami et al. 2009]. In addition, the concept of "thin-slices", which establishes that humans judgments based on first impressions are often correlated with subsequent assertions about people [Ambady and Rosenthal 1992], has been used to study how personality traits and self-exhibition behaviors are displayed in social networks' online profiles, and how they relate to the ways in which users are perceived and interact in their networks [Evans et al. 2008].

In terms of studying nonverbal behavior, our work relates to both classic research in communication theory and a relatively recent series of works in behavioral and ubiquitous computing which have shown that nonverbal cues [Knapp and Hall 2005] are robust and efficient descriptors of human behavior, and that they are consistent indicators of a number of attitudes, attributes, and intentions of people in multiple communication scenarios [Pentland 2008; Gatica-Perez 2009]. In particular, the automatic analysis from "thin-slices" of nonverbal cues from audio and video has proven to be a good estimator of functional roles [Zancanaro et al. 2006] and dominance [Jayagopi et al. 2009] in group meetings, and a reliable predictor of the outcome of salary negotiations [Curhan and Pentland 2007], among other tasks. In contrast to all these works, we analyze a new type of human communication, that is not face-to-face but resembles it in terms of the richness of nonverbal behavior displayed by vloggers and interpreted by a potentially huge audience. Videoblogging has unique feature, including monologue-like content, an asynchronous one-to many nature, and the temporal recurrence of video posts. The similarity with face-to-face interactions comes from the conversational communication intent of vlogging and the fact that users display themselves in the videos. On the other hand, as opposed to the controlled settings, and the high-quality sensors used in most face-to-face behavioral research [Jayagopi et al. 2009; Zancanaro et al. 2006], vlogs result from a widely varying process of video creation and editing. Consequently, vlogs often result in highly diverse content, which poses challenges that need to be addressed with processing techniques that are both robust and applicable at large-scale.

Finally, the remote nature of vlogging draws a link between our work and research on videoconferencing. Human-computer interaction (HCI) and computer-supported collaborative work (CSCW) research has been concerned with issues such as the poor adoption of videoconferencing in favor of face-to-face or phone conversations [de Vasconcelos Filho et al. 2009], the differences in people's behavior between remote scenarios and face-to-face conversations [Sellen 1995], and the effect of image appearance on experience of comfort [de Vasconcelos Filho et al. 2009; Nguyen and Canny 2009]. Though we do not directly address any of these issues, we believe that research in these areas can benefit from each other. For example, in our work, we use visual features that are motivated in one of these works [Nguyen and Canny 2009].

4. THE YOUTUBE DATASET

For this study, we gathered a dataset of conversational vlogs extracted from YouTube. As a preliminary step, we used the YouTube API to automatically query videos using the keywords "vlog", "vlogging", and "vlogger". Based on the manual examination of up to 300 randomly sampled results, we found that

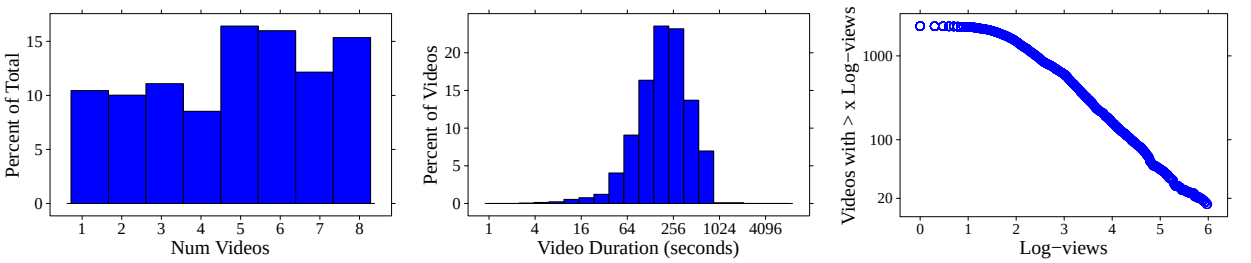

Fig. 3: Some basic figures of the YouTube dataset: (left) the distribution of vlogs collected for the 469 users; (center) the histogram of the videos duration; (right) the cumulative distribution of views per video received in YouTube.

25% of the videos retrieved corresponded to conversational vlogs displaying a single person. The rest of the videos included conversational vlogs displaying more than one person, as well as other types of vlogs such as home and outdoor video footage, music videos, and mashups. Furthermore, by examining the YouTube channels of several users, we observed that the number of conversational vlog entries and their frequency of upload differ substantially among vloggers. Whereas some YouTube users vlog regularly as a core activity, which results in vlog collections of a substantial size, other users vlog to complement a main activity different than conversational vlogging.

As a follow up, we extracted a list of 878 different *usernames* from videos retrieved in November 2009 using the aforementioned keywords. We then recruited 10 untrained volunteers to perform annotations, whose only requirement was to be familiar with YouTube as a video viewer. Aiming to gather conversational vlogs from users' video collections while filtering non-conversational vlogs, we asked the volunteers to annotate up to the most 8 recent videos of non-overlapping samples of users, resulting in one annotation per video. All the videos annotated were publicly available on YouTube. We explicitly recommended annotators to browse the videos using the progress bar, instead of watching them completely, and asked them to answer a few questions about the content. The questions asked to identify the presence of the talking-head setting, the number of unique persons featured in the vlog, and the conversational aspect (as opposed, for example, to the vlogger playing music or singing). Typically, each volunteer spent one hour to annotate the videos corresponding to 25 users. Overall, we obtained annotations for 6,396 videos, as some users had less than 8 videos, which we used to identify a set of *predominantly conversational* vlogs featuring a single person.

Our dataset includes 151 hours of video corresponding to 2,269 videos and 469 different users, as well as the videos' metadata (title, description, duration, keywords, video category, date of upload, number of views, and comments). Figure 3 shows three aspects of our collection. The distribution of conversational vlogs per vlogger in the dataset (Figure 3, left) is slightly skewed towards more than four vlogs ($mean = 4.8, median = 5$). As shown in Figure 3 (center), these vlogs have typical durations between 1 and 6min (70% of the videos appear in this interval), with a median duration of 3min 15s. Only 2.4% of the videos are longer than 10min, a limitation that can be only exceeded by certain users, called partners, who participate in the advertising programs of YouTube. Indeed, this result concurs with the well known tendency of online videos of being short [Cha et al. 2007]. Once individual vlogs are aggregated for each user, this corresponds to over 7min of video per vlogger for 80% of the vloggers in the collection, which represents a large amount of behavioral data compared to typical "thin-slice" sizes. Finally, Figure 3 (right) shows the cumulative distribution of the number of views received by videos. In our sample, the views distribution is skewed towards a small number of views ($median = 231$, $mean = 20030$) with 25% of the videos below 80 views, which is in part a consequence of obtaining the collection from the most recent videos.

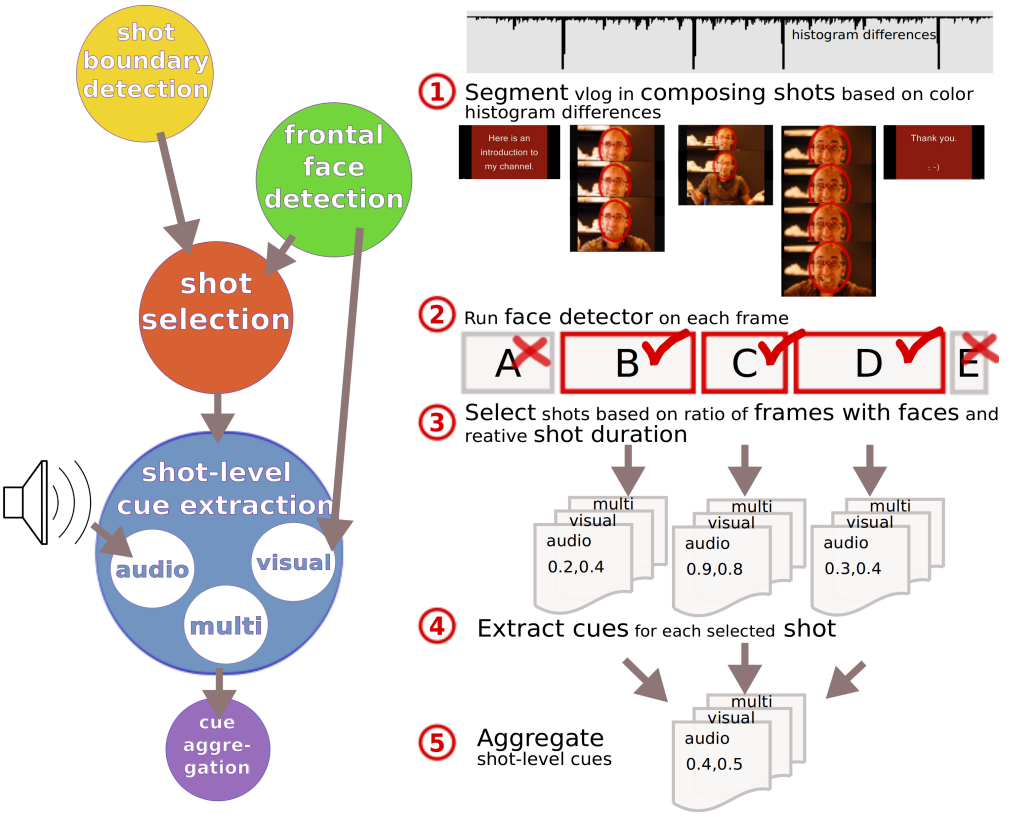

Fig. 4: Automatic processing of vlogs: preprocessing (steps 1, 2, and 3) and nonverbal cue extraction (steps 4 and 5).

5. AUTOMATIC PROCESSING OF VLOGS

As introduced in Section 2, vlogs result in extremely diverse content, compared to purely conversational data recorded in controlled scenarios. For the purpose of analyzing conversational interaction in vlogging, we require audiovisual preprocessing techniques to discard non-conversational content (e.g. openings, closings, or intermediate video snippets containing slideshows or other video footage). In addition, the large-scale feasibility of the analysis calls for robust and computationally efficient processing techniques that can cope with the variety existing in vlogs, both in terms of video quality and behavior in front of the camera.

Our automatic processing approach is illustrated in Figure 4. First, we use a preprocessing scheme to divide vlogs in shots and to identify those shots that display a talking-head. Then, we extract nonverbal cues from conversational shots as descriptors of vloggers' nonverbal behavior. We describe these two components in Sections 5.1 and 5.2, respectively.

5.1 Preprocessing: Conversational Shot Selection

Our preprocessing scheme is grounded on two main assumptions: 1) visual content in conversational shots differs widely from non-conversational shots, and 2) conversational shots can be identified by the presence of a talking-head (or upper body). We use the first assumption, in step (1), to employ a shot boundary detector to segment vlogs in composing shots. Regarding the detection of talking-heads,

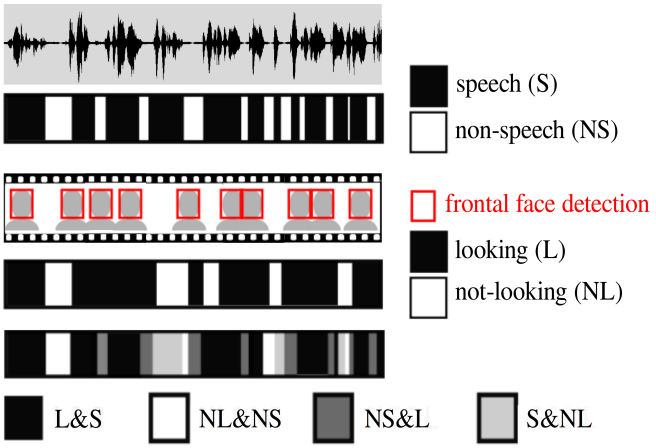

Fig. 5. Nonverbal cues are extracted based on speech/non-speech, looking/non-looking segmentations, and multimodal segmentations.

in step (2), we simplified the task with the detection of frontal faces, a reasonable solution given the inherent nature of conversational vlogging. In addition to its robustness, a face detector may generalize better to the case of vloggers who do not display much of their upper body. For each shot, we assessed the presence of a talking-head by measuring the ratio of frames with face detections. Then, in step (3), we selected conversational shots based on a linear combination of the face detection rate and the duration of the shot relative to the whole duration of the video. This latter condition is motivated by the observation that non-conversational shots tend to be short, independently on whether they feature people or not.

We used existing implementations of algorithms based the OpenCV library [Bradski and Kaehler 2008]. The shot boundary detector finds shot discontinuities by thresholding the Bhattacharyya distance between RGB color histograms of consecutive frames. The face detector implements the boosted classifiers and Haar-like features from the Viola-Jones algorithm [Viola and Jones 2002] using an existing cascaded on the OpenCV version 2.0 [Bradski and Kaehler 2008], which scans faces as small as 20x20 pixels. The shot boundary detection algorithm (step 1) and

and the conversational shot selection (step 3) were tuned in a development set constructed from a small random sample of 100 vlogs. For this purpose, we first annotated hard shot discontinuities (a total of 168 hard shot cuts) and then labeled the shot conversational state ($s = 1$: conversational, $s = 0$: non-conversational). For shot boundary detection, we experimented with different thresholding methodologies, including global, relative, and adaptive thresholds [Hanjalic 2002], and we obtained the best performance ($EER = 15\%$) using a global threshold $\gamma = 0.5$. This performance suggests that the task of boundary detection in conversational vlogs is often simple, and that consecutive composing shots do indeed differ in their content. For conversational shot selection, we first predicted the conversational state s of each shot as $s = \alpha r_f + \beta r_d$, where r_f is the ratio of frames with faces detected, r_d is the duration of the shot relative to the whole vlog duration, and α and β are coefficients obtained using linear regression ($\alpha = 0.76$, $\beta = 0.24$, $R^2 = 0.6, p < 10^{-6}$). Then, we classified shots by thresholding the conversational state s and obtained the best results with a threshold $\gamma = 0.29$ ($EER = 7.5\%$).

5.2 Nonverbal Behavioral Cues Extraction

In this article, we investigate a number of automatic nonverbal cues extracted from both audio and video that have been effective to characterize social constructs related to conversational interaction in both social psychology [Knapp and Hall 2005] and more recently in social computing [Pentland 2008;

Gatica-Perez 2009]. Vocalic and motion cues have shown to be correlated with levels of interest, extraversion, and openness, and are good predictors of dominance [Jayagopi et al. 2009], status [Jayagopi et al. 2008], and the outcome of interactions [Curhan and Pentland 2007]. In addition, we explore multimodal cues, which have also been studied in multi-party conversations [Jayagopi et al. 2009]. Though vlogs are not face-to-face conversations, it is clear that vloggers often behave as if they were having a conversation with their audience, and thus, to some extent, these cues may be suitable to characterize their behavior.

In step (4) (see Figure 5), we extract nonverbal cues for each conversational shot obtained from the preprocessing scheme. Then, in step (5), we aggregate features at the video level. In Sections 5.2.1, 5.2.2, and 5.2.3 we present the list of audio, visual, and multimodal cues we investigated. In addition, in Section 5.2.4, we define a few features that we considered as potentially interesting to explore the categorization of vloggers' at the level of video editing used in their vlogs.

5.2.1 *Audio cue extraction.* We extracted a set of audio cues using the toolbox developed by the Human Dynamics group at the MIT Media Lab, which has proven to be robust in multiple conversational situations [Pentland 2008]. This toolbox implements a two-level hidden Markov model (HMM) to segment the audio in voiced/unvoiced and speech/non-speech regions (see Figure 5), which are used to compute various statistics on speaking activity, fluency, and emphasis as follows:

—*Speaking time (ratio).* Cumulative sum of the speech segment duration divided by the total duration of the video. This is a measure of speaking activity (how much a vlogger talks).

—*Average length of speech segments (in seconds).* Cumulative sum of the speech segment duration divided by the total number of speech segments. This is a measure of speech fluency, which typically relates to the duration and relative number of silent pauses (long segments are associated with short and few pauses) [Scherer 1979].

—*Speaking turns (in Hz).* Total number of speech segments divided by the total duration of the video. This is another measure of speech fluency, which is directly related to the relative number of silent pauses (more speaking turns are associated with more pauses) [Pentland 2008].

—*Voicing rate (in Hz)*. Total number of voicing segments divided by the total duration of the speech segments. This is is another measure of fluency, which relates the frequency of phonemes produced during speech, i.e how fast a person speaks [Scherer 1979].

—*Speaking energy (mean-scaled standard deviation)*. The standard deviation of speech energy divided by the mean speech energy. It is a measure of how well a vlogger controls loudness, which typically relates to emotional stability [Pentland 2008].

—*Pitch variation (mean-scaled standard deviation)*. The standard deviation of pitch divided by the mean pitch. It is a measure of how well a vlogger controls tone, and it is also related to emotional states [Pentland 2008].

5.2.2 *Visual cue extraction.* Relatively few works in conversational modeling have extracted visual activity cues related to hand motion, body motion, and visual focus of attention [Jayagopi et al. 2009]. These techniques usually require good color models, manual initialization, and might fail when used in challenging unconstrained conditions regarding lighting, image quality, resolution, color response, etc. Here, we explore the use of the face detector output to derive coarse measures of gaze and motion, under the sensible assumption that frontal face detections occur when the vlogger looks towards the camera. Though we are clearly not able to estimate the actual direction of the eyes, this method results in a reasonable simplification given the typical vlog setting in conversational shots. On one hand, we use a smooth version of the face detection sequences to construct a looking/non-looking segmentation (see Figure 5), which we employ to compute looking activity features. On the other hand, we use the

position and dimensions of the face detection bounding-boxes to measure the vloggers' proximity to camera and their motion:

—*Looking time (ratio).* Cumulative sum of the looking segment duration divided by the total duration of the video. This measures the looking activity (how much the vlogger looks to the camera).

—*Average length of looking segments (in seconds).* Cumulative sum of the looking segment duration divided by the total number of looking segments. This measures the "persistence" of a vlogger's gaze.

—*Looking turns (in Hz).* Total number of looking segments divided by the total duration of the video. This measures how often a vlogger looks to the camera.

—*Proximity to the camera (ratio).* Cumulative sum of face detection bounding-box size (normalized by the video frame size) divided by the number of frames with detected faces. Larger ratios correspond to shorter distances to the camera. This models the vlogger's choice of addressing the camera from close-ups.

—*Vertical framing (ratio).* Vertical Euclidean distance between the face detection bounding-box center and the frame center (normalized by the video frame height) accumulated and divided by the number of frames with faces. Negative ratios indicate faces positioned towards the upper part of the frame, and are associated with vloggers showing their upper body.

—*Absolute vertical (resp. horizontal) head motion (ratio).* Variation of the absolute Euclidean vertical (resp. horizontal) distance between the bounding-box center and the frame center, given by the standard deviation across frames divided by the mean. Higher values are associated with higher motion.

5.2.3 *Multimodal cues.* The proportion of time spent "looking while speaking" and "looking while listening" to face-to-face conversational partners has been found useful to determine dominance in dyadic conversations [Dovidio and Ellyson 1982] and group meetings [Hung et al. 2008]. Speakers exhibiting the highest "looking while speaking"-"looking while listening" ratios are found to be rated by people as more dominant than those exhibiting a moderated ratio [Dovidio and Ellyson 1982]. Here, we explore a modified version of this ratio which considers the proportion of time "looking while not speaking", to account for the fact that there is only one speaker. First, using the speaking/non-speaking and looking/non-looking segmentations described in Sections 5.2.1 and 5.2.2, we obtain a multimodal segmentation of speaking and looking patterns (see Figure 5). Then, we compute the percentage of time "*looking while speaking*" (L&S), "*looking while not speaking*" (L&NS), and the ratio $L\&S/L\&NS$.

5.2.4 *Video editing cues.* Complementary to nonverbal behavioral cues, we explore the use of the *number of shots per second* and the *video duration* (in seconds) as measures that may reflect vloggers' video editing practices. For example, highly edited videos could be associated with a higher number of shots per second, compared to raw videos recorded in one-take and uploaded without editing.

6. EXPERIMENTS AND RESULTS

As introduced in Section 1, the aim of this article is to address the analysis of vlogger behavior by means of automatic techniques. We divide our analysis in four parts. In Section 6.1, we explore vloggers' practices on video creation. In Section 6.2, we provide an analysis of vlogger behavior based on their aggregated nonverbal cues. In Section 6.3, we study the association between vlogger behavior and the social attention received by their videos. Finally, Section 6.4 discusses the limitations of our analysis.

Table I. : Elements manually coded in a sample of 100 vlogs. The first six elements correspond to video editing elements.

Element	*Description*
Snippets	Opening, ending or intermediate non-conversational video snippets
Opening	Preface video snippet used as a transition to the main conversational part of the vlog
Ending	Credits-like closing video snippet
Intermediate	Snippet that interrupts the conversational scene
Background music	Music used as background during the conversational scene
Sound track	Music used in openings, endings, and intermediate snippets
Object	Vloggers bring an object to the camera around which the conversation takes place

6.1 Vloggers' Video Creation Practices

With the purpose of identifying some of the vloggers' common practices in terms of video composition and editing, we performed a manual content analysis of a small sample of vlogs from our dataset. Since this type of analysis is not feasible for the whole dataset, in a second step, we exploited the output of both the shot segmentation and conversation selection methods to bring up several insights on the large-scale patterns of vloggers.

Our manual content analysis of a random sample of 100 vlogs is similar to the one conducted by Landry and Guzdial [2008]. However, compared to them, we focus on conversational vlogs as a particular type of online video, bringing insights on this specific type of content. We analyzed our sample using the codes shown in Table I. For each vlog, we coded elements as being present or absent in the vlog (independently of how many times they occurred). This was done because we are not interested on the frequency of the elements in the sample, but on their popularity in vlogging as given by the percentage of vlogs using them.

Figure 6 shows the percentage of videos featuring each one of the video editing elements that were manually coded. Overall, the most popular practice among vloggers is the introduction of video snippets during video editing. 45% of the vlogs contained some kind of non-conversational snippet together with the main conversational scene. Most commonly, these video snippets interrupted the main conversational scene, as evidenced by 32% of the videos containing at least one intermediate non-conversational snippet. As observed during the coding process, these type of snippets typically correspond to video footage or edited slideshow sequences of portraits, landscapes, and text, which displayed people or events mentioned during the conversation. Similarly, when referring to them, vloggers usually brought the objects (e.g. books, dvds, electronics) towards the camera, or moved the camera towards them (e.g. the vlogger shows the room from where he vlogs). This behavior, found in 26% of the vlogs, reveals the communicative intention of some vloggers, which develops partly or totally around these elements. Openings and endings appeared in 16% and 20% of the vlogs, respectively. Compared to intermediate snippets, openings and endings typically used text and images only, to create prefaces and closings that are repeated along the vloggers' collection as a "branding" element. Finally, in terms of audio, we found that 25% of vlogs used some kind of soundtrack in their non-conversational snippets, and to a lesser extent (12% the vlogs), included background music during the whole conversation.

Interestingly, the use of the video editing elements coded in conversational vlogs compares poorly with the use of the same elements in the random sample of the 100 top popular videos analyzed by Landry and Guzdial [2008]. Compared to them, the percentages of videos in our sample containing openings, closings, and soundtracks are about half. This suggests that vloggers focus mainly on the conversational aspect of the video, and exploit video editing to enrich and support their discourse.

Moving beyond manual coding, and as a by-product of the automatic vlog preprocessing, we draw basic statistics about the number of selected (conversational) and rejected (non-conversational) shots

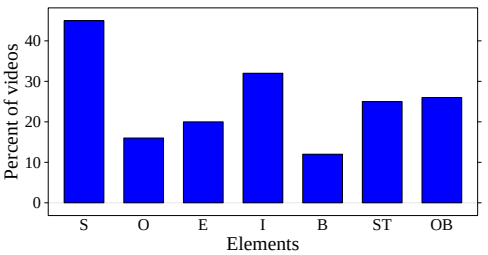

Fig. 6. Common elements coded in a random sample of 100 vlogs: Snippets (S), Openings (O), Endings (E), Intermediate snippets (I), background music (B), Soundtrack (ST), and objects (OB). The Y axis shows the percentage of videos that contained at least one occurrence of the coded element. Videos were coded with S whenever they contained any of the snippets E, O, or I. Thus, the percentage of videos coded with S corresponds to the union of videos coded with O, E, and I.

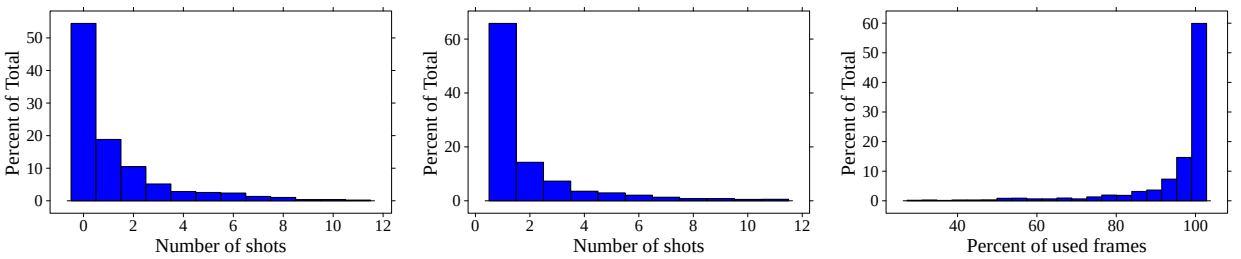

Fig. 7: Distributions of non-conversational (rejected) shots, conversational (selected) shots, and percentage of frames per vlog left after automatic preprocessing.

with the purpose to illustrate some characteristics of the whole dataset. Figure 7 (left) shows the distribution of non-conversational shots per vlog found automatically in the whole dataset. Around 45% of the 2269 vlogs have one or more non-conversational shots which coincides with the percentage of vlogs that were manually coded as containing non-conversational video snippets in the previous section. This implies that more than half of the vlogs consist of monologues (with zero non-conversational shots). Though these vlogs could have been edited from different conversational scenes, the distribution of conversational shots in Figure 7 (center), which provides a complementary view of the problem, indicates that more than 60% of the vlogs do consist of a single conversational shot. This suggests that a large proportion of vloggers shot their vlogs in one take, which may affect the spontaneity of the resulting behavior, or alternatively, that vloggers chose the uploaded take from several recorded ones. Figure 7 (right) shows the percentage of frames per vlog left after removing non-conversational shots, which indicates that non-conversational shots tend to be a small fraction of the content of the video. The numbers shown here concur with the notion that non-conversational shots are used as elements of support to the core conversational part of the video.

Overall, the results in this section provide insights about the nature of conversational vlogging: people often vlogcast themselves without much editing, and when they use editing, it is used judiciously. Furthermore, the editing itself can be robustly detected by our processing techniques, which allows the extraction of the actual conversational segments of vlogs, which we analyze in the following section.

6.2 Vloggers' Nonverbal Behavior

We examined the automatic extraction of nonverbal cues by plotting their distributions and computing some basic statistics, both at shot level and video level. We select some of the nonverbal cues and describe relevant aspects of their histograms in the context of conversational scenarios. In addition, we provide a correlation analysis between cues intra and inter-modality.

Figure 8 shows the distribution of some nonverbal cues obtained at the shot-level. After aggregating the features for each video, these distributions show smoother tails but overall little differences, which

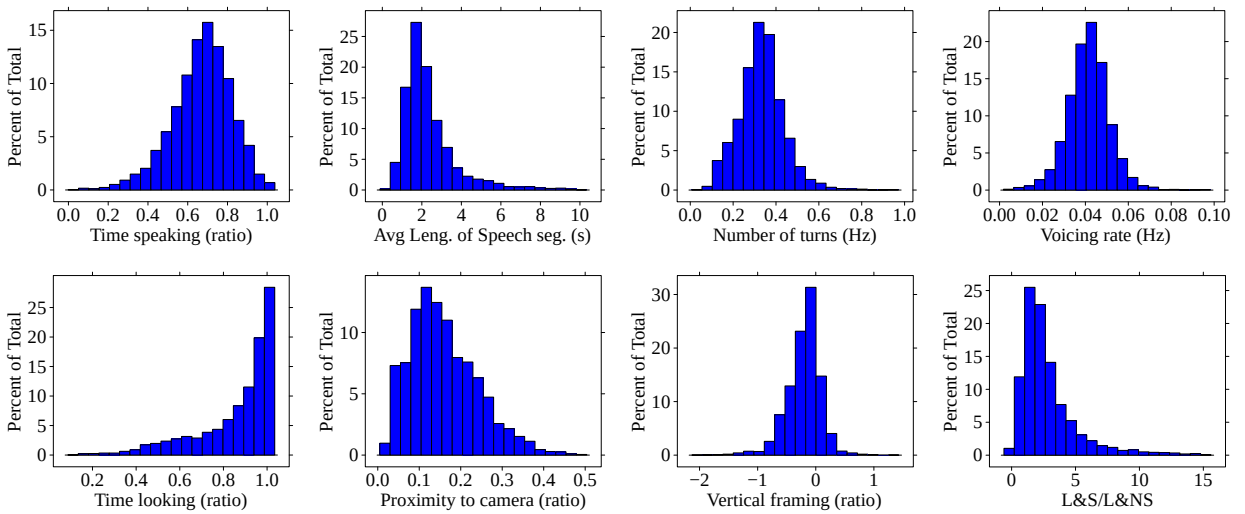

Fig. 8: Selected nonverbal cue distributions for conversational shots in YouTube vlogs: four audio cues, three visual cues, and one multimodal.

result from the fact that most of the vlogs are composed of few conversational shots (see Section 6.1). These distributions unveil information that may be useful to understand some basic characteristics of nonverbal behavior in vlogging. For example, the speaking time distribution, biased towards high speaking times ($median = 0.65$, $mean = 0.67$, $sd = 0.15$), shows that 85% of the conversational shots contain speech more than half of the time, which suggests that vloggers who were perceived as mainly talking during the annotation process (Section 4) are indeed speaking for a significant proportion of the time. Speaking segments tend to be short ($median = 1.98s$, $mean = 2.36s$, $sd = 1.36s$), which is common in spontaneous speech, typically characterized by higher numbers of hesitations and lower fluency [Levelt 1989]. The median number of speaking turns per second ($median = mean = 0.33$, $sd = 0.10$), which corresponds to one speaking turn every 3 seconds, evidences that pauses between speaking segments are also short. Finally, the voicing rate ($median = mean = 0.4$, $sd = 0.09$) varies between 2 and 4 regions per second, a range of values similar to other conversational scenarios [Scherer 1979].

Regarding visual cues, the looking time ($median = 0.68, mean = 0.67, sd = 0.14$) is biased towards high values, with 50% of the vloggers looking at the camera over 90% of the time. It is not entirely clear to what extent this corresponds to a pure "addressing the camera" behavior, rather than the result of the simplification made by assuming that frontal face detections imply that vloggers look at camera. Considering the typical frame size of a YouTube video in our dataset (320x240 pixels) and the proximity to the camera ($median = 0.19, mean = 0.23, sd = 0.14$), vloggers' face size varied approximately between 19x15 and 176x132 pixels, with a median of 64x48 pixels. Since smaller ratios indicate larger distance to the camera, these figures suggest that vloggers typically respect some "standard" distance to the camera, neither being too close nor too far. In addition, as shown by the vertical framing ratios ($median = -0.17, mean = -0.21, sd = 0.29$), faces are typically positioned in the top half of the frame, which is associated with vloggers showing their upper body.

Regarding multimodal cues, the ratio L&S/L&NS ($median = 2.25, mean = 3.87, sd = 12.79$) shows that vloggers tend to look at the camera when they speak more frequently than when they are silent.

Table II. : Pearson's intra-feature correlations /$^{*}p < .01$,** $p < .001$,*** $p < .0001$).

		1	2	3	4	5	6	7	8	9	10	11
1	Speaking time	–	.74***	−.53***	.23***	.14***	−.16***	−.00	−.03	−.11***	.69***	−.72***
2	Av Len Speak Seg		–	−.84***	.06	.05	−.05	−.01	−.02	−.03	.47***	−.62***
3	# Speech Turns			–	.03	−.02	−.01	.00	−.01	−.01	−.30***	.52***
4	Looking time				–	.51***	−.85***	−.04	.02	−.31***	.76***	.33***
5	Av Len Look Seg					–	.50***	−.04	.04	−.21***	.43***	.15***
6	# Look Turns						–	.04	−.01	.35***	−.68***	−.33***
7	Proximity to Cam							–	.38***	−.26***	−.03	−.02
8	Ver Frame								–	−.18***	−.02	.04
9	Ver Head Motion									–	−.27***	−.09***
10	L&S										–	−.16***
11	L&NS											–

This resembles the behavior of dominant people in dyadic conversation, who tend to look at others more while speaking than while listening [Dovidio and Ellyson 1982].

Finally, we computed Pearson's correlations between the average nonverbal cues for pairs of features of all the modalities and we summarized them in Table II. Some nonverbal cues extracted from audio and video show moderate and large correlations within the same modality. For example, the speaking time is positively correlated to the average length of speaking segments ($r = .75$, $p < 10^{-3}$) and is negatively correlated to the speaking turns ($r = -.53$, $p < 10^{-3}$). Similarly, the looking time is positively correlated to the average length of looking segments ($r = .51$, $p < 10^{-3}$), and negatively correlated to the looking turns ($r = -.85$, $p < 10^{-3}$). Interestingly, the motion is correlated negatively with the proximity to the camera ($r = -.26$, $p < 10^{-3}$), reflecting that being close to the camera allows for less activity if the speaker is supposed to be framed in the video. The correlations between audio and visual nonverbal cues are lower, as for example, between the speaking time and the looking time ($r = .23$, $p < 10^{-3}$). In addition, the multimodal cues are significantly correlated with both patterns of speaking and looking. See for example, the correlation between "looking while speaking" and speaking time ($r = .69$, $p < 10^{-3}$) or between "looking while speaking" and looking time ($r = .76$, $p < 10^{-3}$). Although some of these correlations may seem low, overall, these they are withing the levels often reported in social psychology [Dovidio and Ellyson 1982; Ambady and Rosenthal 1992].

6.3 Vloggers' Nonverbal Behavior and Received Social Attention

The analysis of individual correlates is a standard approach to the study of nonverbal behavior in social psychology research [Scherer 1979; Dovidio and Ellyson 1982; Ambady and Rosenthal 1992], and thus, it is adequate to address the study of vloggers' nonverbal behavior. In social computing, works have shown that certain nonverbal cues correlate with several perceived social attributes (e.g. dominance, role, attraction) during face-to-face interactions [Pentland 2008; Gatica-Perez 2009]. In our work, we hypothesize that similar effects may take place in conversational vlogs. Typically, these nonverbal cues emerge naturally from the behavior of certain personalities, attitudes, or skills of people who, in some manner, are successful in their communication exchanges. We hypothesize that nonverbal behavioral cues correlate with the social attention to vlogs, a measured derived from the number of views they receive. As suggested in a recent work [Huberman et al. 2009], the attention received is a valuable resource in social media that can be understood as a public reward to users who contribute with quality content. This idea concurred with the finding that YouTube users' productivity exhibits a strong positive dependence on attention, and that a lack of attention leads to a decrease in the number of videos uploaded [Huberman et al. 2009]. Compared to the phenomenon of popularity, broadly studied in social media, the average level of attention is a coarser measure that may be useful to explain

broader patterns in terms of *subpopulations* of vloggers who behave similarly, i.e. who tend to speak for the same amount of time or look at the camera in similar ways. In a recent paper [Biel and Gatica-Perez 2010], we explored the correlation between nonverbal behavior and popularity by measuring the pair-wise correlations between nonverbal cues extracted from audio and the exact number of vlogs' views. The experiments, which resulted into moderate correlations, were in general dominated by the large dispersion of videos' views, given by their power-law distribution.

Here, we use the following method to analyze the association between nonverbal behavior and social attention. We define the *average level of attention* ($\hat{v}$) of a set of N videos as $\hat{v} = median\{\log v_n\}_{n=1}^{N}$, where v_n is the number of views received by the n-th video. For each nonverbal cue, we divide the set of all videos into L approximately equally sized groups corresponding to L different nonverbal cue levels, and define the *average nonverbal cue* for the l-th level as $\hat{c}_l = mean\{c_{n,l}\}_{n=1}^{N_l}$, $l = 1 \dots L$, where $c_{n,l}$ is the nonverbal cue computed for the n-th video in the level l, and N_l is the number of videos in that level. Based on this grouping, we then compute Pearson's correlation between the average nonverbal cue distribution $\hat{c}_l, l = 1 \dots L$ and the corresponding average level of attention $\hat{v}_l, l = 1 \dots L$, computed as defined above.

Table III shows the correlations between selected nonverbal cues (from audio, visual, and multimodal cues) and social attention, measured using $L = 50$ levels. For each video, nonverbal cues were aggregated from shots using different methods: taking the mean or median, taking the mean after weighting the cues proportionally to the shot duration, or taking the values from the longest shot. We observed that different aggregation methods produced quite similar results, partly because of a large proportion of videos containing only one or two shots. To further illustrate the relation between nonverbal cues, Figure 9 shows the xyplot of some of these cues with the median number of log-views. One could argue that this analysis of correlations is only valid if the distributions of views for the levels are significantly different. To test this condition, we conducted a Welch's test of the null hypothesis H0: "The distributions of the levels 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 levels between 10 and 100 and obtained p-values lower than 0.001, which suggests that the hypothesis can be rejected.

Regarding audio cues, the speaking time, the average length of speech segments, and the number of turns are the features showing a larger correlation with social attention (up to $r = .84$, $r = .86$ and $r = -.72$, $p < 10^{-3}$). Interestingly, speaking activity has been reported in social psychology works as being among the most effective nonverbal cues to predict social constructs such as dominance, or physical attractiveness of participants in conversational scenarios [Knapp and Hall 2005; Pentland 2008]. In the case of vlogging, the results indicate that vloggers talking longer, faster, and using fewer pauses receive more views from their audiences. The voicing rate and the variation of energy show smaller yet significant correlations (up to $r = .26$, $r = -.32$, $p < 10^{-2}$), which suggests that speaking faster, and having vocal control might be also related to the way vloggers are perceived in YouTube. These results compare to findings in face-to-face interactions, where, for example, these specific cues were predictors of success on salary negotiations [Curhan and Pentland 2007].

Several visual cues are also significantly correlated with social attention. Similarly to speaking patterns, looking patterns are largely correlated with the median number of log-views (up to $r = .70$, $p < 10^{-4}$). In the same way, the time looking at the camera and the average duration of looking turns are positively correlated with attention (up to $r = .48$ and $r = .70$, $p < 10^{-4}$), whereas the number of looking turns shows a negative correlation. Perhaps one of the most interesting results is the negative correlation between the vloggers' proximity to the camera and the median number of log-views (up to $r = --62$, $p < 10^{-4}$), which suggests that respecting an "optimal" distance to the camera might have an effect on the communication process of vlogging, somehow penalizing those vloggers being too close to the camera. Though this result may not be obvious, a recent study on the effect of video framing

Table III. : Pearson's correlation between nonverbal cues and median number of log-views, for different aggregation methods ($^{*}p < .01, ^{**} p < .001, ^{***} p < .0001$, m-sd= mean-scaled standard deviation)

Features	Shot aggregation method			
Audio cues	Median	Mean	Weight	Longest
Speaking time	$.81^{***}$	$.82^{***}$	$.84^{***}$	$.80^{***}$
Av Len Speak Seg	$.80^{***}$	$.79^{***}$	$.80^{***}$	$.86^{***}$
# Speech Turns	$-.69^{***}$	$-.64^{***}$	$-.70^{***}$	$-.72^{***}$
Voice rate	.23	.20	$.26^{*}$	.21
Speaking energy (m-sd)	$-.26^{*}$	$-.30^{*}$	$-.32^{*}$	$-.21$
Pitch (m-sd)	.10	.09	.12	.06
Visual cues	Median	Mean	Weight	Longest
Looking time	$.62^{***}$	$.53^{***}$	$.70^{***}$	$.50^{***}$
Av Len Look Seg	.19	$.29^{*}$	$.48^{***}$	$.48^{***}$
# Look Turns	$.62^{***}$	$.53^{***}$	$.70^{***}$	$.50^{***}$
Proximity to Camera	$-.62^{***}$	$-.61^{***}$	$-.57^{***}$	$-.60^{***}$
Ver Frame	$-.83^{***}$	$-.84^{***}$	$-.84^{***}$	$-.85^{***}$
Hor Head Motion	$.48^{***}$	$.70^{***}$	$.69^{***}$	$.69^{***}$
Ver Head Motion	$.31^{***}$	$.40^{***}$	$.49^{***}$	$.52^{***}$
Multimodal cues	Median	Mean	Weight	Longest
L&S	$.75^{***}$	$.79^{***}$	$.76^{***}$	$.73^{***}$
L&NS	$-.78^{***}$	$-.73^{***}$	$-.74^{***}$	$-.80^{***}$
L&S/L&NS	$.69^{***}$	$.68^{***}$	$.73^{***}$	$.68^{***}$

(i.e. the proportion of the human body that appears within the frame) on the empathy of participants in videoconferences found significant differences between head-only and upper-body framing [Nguyen and Canny 2009]. In fact, the proportion of video frame occupied by the face could be thought as a proxy for the amount of body shown to the camera, in a way that the smaller the face, the larger the proportion of upper-body shown, and so the higher the possibility for the audience of perceiving body language cues. The proportion of upper-body shown is likely better measured by the vertical framing, which in our results shows a larger correlation with social attention with the face size. The two measures of motion revealed a significant positive correlation with attention (up to $r = .70$, $r = .52$, $p < 10^{-3}$). Interestingly, other works have suggested that successful people in meeting interactions tend to be more visually active [Gatica-Perez 2009].

The multimodal cues "looking-while-speaking" (L&S) , "looking-while-not-speaking" (L&NS), and the ratio L&S/L&NS also show significant correlations (up to $r = .79$, $r = .80$, $r = .73$, $p < 10^{-3}$ respectively) for the median. Multimodal cues based on speaking and looking turns have also been found to be effective in predicting dominance in multi-party conversations [Jayagopi et al. 2009].

Finally, the measures of video editing proposed such as the number of shots and the video duration show low and no significant correlation with social attention, respectively ($r = .35$, $p > 10^{-2}$ and $r = .08$, $p > 0.1$ respectively). It is not clear to what extent these results follow from the effectiveness of these two features to capture the level of complexity of video editing.

6.3.1 *Accounting for the Temporal Dimension of Videos' Views.* So far in our analysis, we used the accumulated view count of YouTube vlogs to measure social attention, without considering the date of upload of the video. However, one could argue that older videos could accumulate a larger number of views than recently uploaded videos, as they have been publicly exposed for a longer time. To measure this effect, we divided the videos in $G = 50$ approximately equally-sized groups corresponding to different dates of upload. For each group, we computed the videos' average age $\hat{\tau}_g = mean\{\tau_n\}_{n=1}^{N_g}$, where

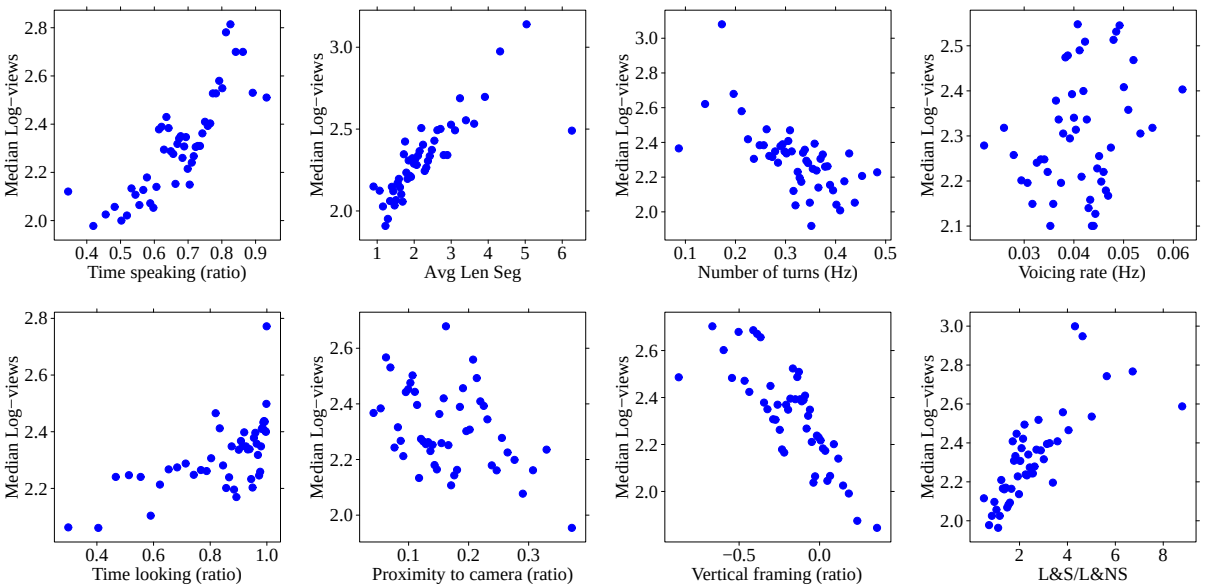

Fig. 9: The aggregated nonverbal cues versus the social attention measure.

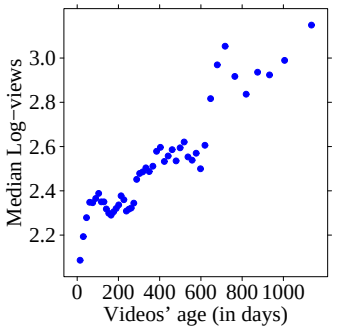

Fig. 10. Social attention vs. age of videos (number of days between videos' upload date and data collection date). The average level of attention increases as the age of the video increases with very high correlation ($r = .94$, $p < 10^{-3}$).

τ_n is the age of the n-th video in the group computed as the number of days between the video's upload date and the data collection date, and N_g is the number of videos in the group. We then computed Pearson's correlation between the videos' average age $\hat{a}_g$, $g = 1 \dots N$, and the average level of attention $\hat{v}_g$, $g = 1 \dots N$. As shown in Figure 10, this correlation is strong ($r = .94$, $p < 10^{-3}$). To compensate for this effect, we used a linear regression model to predict the average level of attention of videos as a function of the videos' average age: $\hat{pv}_g = a\tau + b$, where $\hat{pv}_g$ is measured in log-views, τ is the video average age, and a and b are the coefficients obtained by linear regression ($a = 8 \times 10^{-4}$, $b = 2.19$, $R^2 = 0.88$, $p < 10^{-3}$). With this model, we updated the original views log-count of each video in the data set to $\log vd_i = \log v_i - \hat{pv}_g$, where vd_i are the updated views, and $\hat{pv}_g$ is the predicted average level of attention for the age group of the i-th video. We then replicated all the experiments discussed in Section 6.3 computing the social attention based on the updated views, i.e. $\hat{v} = median\{\log vd_n\}_{n=1}^{N}$. As in the experiments carried out by Huberman et al. [2009], accounting for the temporal dimension of the views did not result in significant differences on the correlation effects measured. We argue that this results from the aggregation-based method used for the analysis, which benefits from grouping

several videos in a single group to account for the the dispersion of views distribution among different videos, independently of their age.

6.4 Limitations of our study

The automatic processing of vlogs, as addressed in this article, has several limitations. First, the shot boundary detection and the conversational shot selection of the preprocessing step were optimized on a sample of 100 vlogs, which may seem small compared to the whole dataset. However, for this set, we manually annotated 168 shot boundaries, a number that compares well with datasets used in earlier works specifically addressing the shot boundary detection task [Hanjalic 2002]. Clearly, analyzing the accuracy of the shot boundary detection is prohibitively expensive for the 150h of video in our dataset. In our experiments, the shot boundary detection automatically found 3,278 shot boundaries. In addition, note that the data used to evaluate the conversational shot selection requires annotating shot boundaries first.

Second, the automatic extraction of nonverbal behavior has a main limitation, namely cue validity, in other words, the assessment that each nonverbal cue is appropriately capturing the aspect of conversational dynamics it is supposed to. In general, measuring nonverbal cues manually (also known as coding in the psychology literature) is a much more laborious and expensive task than shot boundary or conversational shot selection. It requires a substantial amount of time and expertise, and thus, it is typically done in social psychology works with small samples of data [Knapp and Hall 2005]. On the other hand, research in social computing has shown that the use of simple, although imperfectly extracted, nonverbal cues are effective to characterize human behavior [Pentland 2008; Gatica-Perez 2009], despite the fact that the accuracy of the automatic nonverbal measures used in these works is not validated for the reason explained above. As discussed by Curhan and Pentland [2007], it should be noticed that the practical impossibility of validating nonverbal cues at large-scale is compensated with high test-retest validity of the feature extraction methods, (i.e. multiple runs of the same feature extraction algorithm in one video provide the same feature values), which is a relevant and desired requirement for behavioral measures in social psychology research.

7. CONCLUSIONS

In this article, we introduced a new research domain in social media, namely the automatic analysis of conversational vlogs, which are multimodal in nature and have a significant potential for large-scale analysis. The goal of this domain is to study the behavior of vloggers based not only on contextual information from metadata, but on the specific ways vloggers themselves behave in the videos, thus exploiting the "vlogcast yourself" metaphor for automatic analysis.

We presented a study of the characterization of vloggers' nonverbal behavior, based on the use of automatic audio and visual techniques that are robust to the variability of content found in this type of videos. The proposed methods included discarding the non-conversational parts of vlogs, and processing the conversational ones to extract descriptors of vloggers' behavior. In particular, we extracted nonverbal cues that have strong support in social psychology research as characterizing social aspects of human communication beyond the spoken words. Our analysis on 2,269 vlogs extracted from YouTube brought novel insights on common practices of vloggers in terms of video creation and editing, as compared to other types of online video. In particular, the automatic processing of conversational vlogs confirm the idea that content in this type of videos is mainly driven by a communicative intent, which may be optionally accompanied by non-conversational snippets.

A more compelling result of our work is the evidence that some audio, visual, and multimodal cues extracted from vloggers' behavior such as the speaking time, the looking time, the proximity to the camera, and the proportion of time "looking-while-speaking" to "looking-while-not-speaking", are cor-

related with the average level of attention of their vlogs. To our knowledge, this is the first time such a connection is investigated. We shall also emphasize that we do not claim in any case any direct causality effect between nonverbal cues and social attention. Rather, thse results may 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. These nonverbal cues are likely related to social constructs such as specific personality traits (like extraversion) or persuasion, and to how effective people are at creating vlogs. We intend to dedicate future work to investigate these dimensions. We would also like to start addressing some aspects of verbal behavior in vlogging (i.e. what they say), which to our knowledge has not been studied yet.

In our view, the goal of behavioral research in vlogs is to obtain a rich, robust, multimodal characterization of vloggers that can ultimately be useful to build automated tools for discovery and interaction with and through vlog content. For example, behavioral analysis can be useful to automatically predict specific social constructs, with applications for vlogger discovery. The automatic analysis of vlogs can also be useful for vlog retrieval, buzz-tracking, opinion mining, and sentiment analysis from vlog data. In addition, one could envision tools that specifically help vloggers to improve some aspects of their content related to human communication and social attention.

ACKNOWLEDGMENTS

This research has been supported by the Swiss National Center of Competence (NCCR) on Interactive Multimodal Information Management (IM)2. We thank the volunteer annotators for their time, as well as the YouTube vlogger community for publicly sharing their creativity through their vlogs. We also thank our colleagues for posing as vloggers.

REFERENCES

AMBADY, N. AND ROSENTHAL, R. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin 111,* 2, 256–274.

BIEL, J.-I. AND GATICA-PEREZ, D. 2009. Wearing a YouTube hat: directors, comedians, gurus, and user aggregated behavior. In *Proc. of the 17th ACM Int. Conf. on Multimedia (MM)*. Beijing, China.

BIEL, J.-I. AND GATICA-PEREZ, D. 2010. Voices of Vlogging. In *Proc. of Int. AAAI Conf. of Weblogs and Social Media (ICWSM)*. Washington, DC.

BRADSKI, G. AND KAEHLER, A. 2008. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly.

BURGESS, J. AND GREEN, J. 2009. *YouTube: online video and participatory culture*. Polity, Cambridge, UK.

CHA, M., KWAK, H., RODRIGUEZ, R., AHN, Y.-Y., AND MOON, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. of ACM SIGCOMM Internet Measurement Conf. (IMC)*. San Diego, CA.

CHENG, X., DALE, C., AND LIU, J. 2008. Statistics and social network of youtube videos. In *Proc. of IEEE 16th Int. Workshop on Quality of Service (IWoQS)*. Enschede, Netherlands.

CURHAN, J. R. AND PENTLAND, A. 2007. Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology 92,* 3.

DE VASCONCELOS FILHO, J. E., INKPEN, K. M., AND CZERWINSKI, M. 2009. Image, appearance and vanity in the use of media spaces and video conf. systems. In *Proc. of ACM Group 2009 Conf.* Sanibel Island, FL.

DOVIDIO, J. F. AND ELLYSON, S. L. 1982. Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Journal of Social and Personal Relationships 45,* 2, 106–113.

EVANS, D. C., GOSLING, S. D., AND CARROLL, A. 2008. What elements of an online social networking profile predict target-rater agreement in personality impressions. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*. Seattle, WA.

GATICA-PEREZ, D. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing 27,* 12, 1775–1787.

GILL, J. A., NOWSON, S., AND OBERLANDER, J. 2009. What are they blogging about? Personality, topic, and motivation in blogs. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*. San Jose, CA.

GOSWAMI, S., SARKAR, S., AND RUSTAGI, M. 2009. Stylometric analysis of bloggers' age and gender. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*. San Jose, CA.

GRIFFITH, M. 2007. Looking for you: An analysis of video blogs. In *Annual Meeting of the Association for Education in Journalism and Mass Communication*. Washington, DC.

HALVEY, M. AND KEANE, M. 2007. Exploring social dynamics in online media sharing. In *Proc. of the 16th Int. Conf. on World Wide Web (WWW)*. Banff, Alberta, Canada.

HANJALIC, A. 2002. Shot-boundary detection: unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology 12,* 2, 90–105.

HARLEY, D. AND FITZPATRICK, G. 2009. Youtube and intergenerational communication: the case of geriatric1927. *Universal Access in the Information Society 8,* 1, 5–20.

HUBERMAN, B. A., ROMERO, D. M., AND WU, F. 2009. Crowdsourcing, attention and productivity. *Journal of Information Science 35,* 6.

HUNG, H., JAYAGOPI, D. B., BA, S., ODOBEZ, J.-M., AND GATICA-PEREZ, D. 2008. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proc. of the 10th Int. Conf. in Multimodal Interfaces (ICMI)*. Chania, Crete, Greece.

JAYAGOPI, D. B., BA, S., ODOBEZ, J.-M., AND GATICA-PEREZ, D. 2008. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proc. of the 10th Int. Conf. in Multimodal Interfaces (ICMI)*. Chania, Crete, Greece.

JAYAGOPI, D. B., HUNG, H., YEO, C., AND GATICA-PEREZ, D. 2009. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing 17,* 3, 501–513.

KNAPP, M. L. AND HALL, J. 2005. *Nonverbal communication in human interaction*. Holt, Rinehart and Winston, New York.

KRAMER, A. D. I. AND RODDEN, K. 2008. Word usage and posting behaviors: modeling blogs with unobtrusive data collection methods. In *Proc. of the 26th annual SIGCHI Conf. on Human factors in computing systems (CHI)*. Florence, Italy.

KRUITBOSCH, G. AND NACK, F. 2008. Broadcast yourself on YouTube - really? In *Proc. of the 3rd ACM Int. Workshop on Human-Centered Computing (HCC)*. Vancouver, British Columbia.

LANDRY, B. AND GUZDIAL, M. 2008. Art or circus? characterizing user-created video on YouTube. Tech. rep., Georgia Institute of Technology.

LANGE, P. 2007. Publicly private and privately public: social networking on youtube. *Journal of Computer-Mediated Communication 1,* 13.

LEVELT, W. J. M. 1989. *Speaking: From Intention to Articulation*. MIT Press, Cambridge, Mass.

LIN, W.-H. AND HAUPTMANN, A. 2008. Identifying ideological perspectives of web videos using folksonomies. In *Proc. of AAAI 2008 Fall Symposium on Multimedia Information Extraction.* Arlington, Virginia.

MISHNE, G. 2005. Experiments with mood classification in blog posts. In *Proc. of SIGIR Stylistic Analysis Of Text For Information Access*. Salvador, Bahia, Brazil.

MISLOVE, A., MARCON, M., GUMMADI, K., DRUSCHEL, P., AND BHATTACHARJEE, B. 2007. Measurement and analysis of online social networks. In *Proc. of ACM SIGCOMM Internet Measurement Conf. (IMC)*. San Diego, CA.

MOLYNEAUX, H., O'DONNELL, S., GIBSON, K., AND SINGER, J. 2008. Exploring the gender divide on YouTube: An analysis of the creation and reception of vlogs. *American Communication Journal 10,* 2.

NGUYEN, D. T. AND CANNY, J. 2009. More than face-to-face: empathy effects of video framing. In *Proc. of the 27th annual SIGCHI Conf. on Human factors in computing systems (CHI)*. Boston, MA.

PENTLAND, A. S. 2008. *Honest Signals: How They Shape Our World*. MIT Press Books Series, vol. 1. The MIT Press.

SCHERER, K. R. 1979. Personality markers in speech. In *Social markers in speech*, K. R. Scherer and H. Giles, Eds. Cambridge: Cambridge University Press, 147–209.

SELLEN, A. J. 1995. Remote conversations: The effects of mediating talk with technology. *Human Computer Interaction 10*, 401–444.

STRANGELOVE, M. 2010. *Watching YouTube: Extraordinary Videos by Ordinary People*. University of Toronto Press.

VIOLA, P. AND JONES, M. 2002. Robust real-time object detection. *Int. Journal of Computer Vision 57,* 2.

VONDERAU, P. 2010. *The YouTube Reader*. Wallflower Pr.

ZANCANARO, M., LEPRI, B., AND PIANESI, F. 2006. Automatic detection of group functional roles in face to face interactions. In *Proc. of the 8th Int. Conf. on Multimodal Interfaces (ICMI)*. Banff, Alberta, Canada.

ZHANG, X., XU, C., CHENG, J., LU, H., AND MA, S. 2009. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia 11,* 2.