MINING CONVERSATIONAL SOCIAL VIDEO

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse présenté le 2013
à la Faculté des Génie Electrique
programme doctoral en Génie Electrique
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Joan-Isaac Biel

jury:

Prof Sabine Susstrunk président du jury
Dr Daniel Gatica-Perez, directeur de thèse
Prof Pascal Frossard, rapporteur
Dr Bernardo Huberman, rapporteur
Dr Mathias Mehl, rapporteur

Lausanne, EPFL, 2013

I have no special talent.
I am only passionately curious.

– Albert Einstein

Acknowledgements

When it comes to life the critical thing is whether
you take things for granted or
you take them with gratitude.

– Gilbert K. Chesterton

On Thursday May 16th, while I was going back home, right after my private defense, I found myself invaded by a sudden sense of gratitude towards many people that have been part of this long journey and that deserve my most sincere recognition. This is my opportunity to acknowledge them by their names.

I'd like to start with a big thank you to Daniel Gatica-Perez, who has worn the hat of director, guru, and friend. He has the gifts of mentorship and innovation. I am grateful for his attention and his patience, and for teaching me how to think out-of-the-box. No doubt a large part of the success of this thesis is due to him. Thank you for these four years of work together.

Thank you to my parents, Francisco Javier Biel and Maria Auxili Tres. Thank you to my dearest sisters Anna, Marta, Roser, Gemma, Mònica, and Leyre, and my brother Xavier. It's no exaggeration if I say that I have never found a comparable example of creativity, hard work, and kindness all together. You have been, and will remain, a great inspiration for me. You know I love you, though I have never been the best at expressing so.

Thank you to my "stepsisters", Mariona López and Judit Fosses, for their love and support. These girls have been my anchor to the outside world for more than three years, and have assisted me in many more ways that I can count. To them, I can only respond with more love.

Thank you to the scientific committee, Prof. Pascal Frossard, Dr. Bernardo Huberman, Dr. Mathias Mehl, and Prof. Sabine Susstrunk, who accepted being part of my jury and provided useful feedback to improve this dissertation. Especially thank you to Mathias for his fresh ideas and his enthusiasm, and to Bernardo for being the only one to tell me how good I looked in my new suit.

Thank you to my colleagues at the social computing group, from whom I have learned so much. Thank you to Edgar Roman (Paco), Dayra Sanchez, Oya Aran, Gokul Chittaranjan, Dinesh Jayagopi, Radu Negoescu, Kate Farrahi, Hari Parthasarathi, Hayley Hung, Laurent Nguyen, Darshan Santani, and Alvaro Marcos. Thank you to those that contributed to make of

305 the best office at Idiap. Thank you to Radu for not killing me on my first day at Idiap and to Paco for still being a friend after four years of sitting in front of each other.
Thank you to my scientific collaborators Oya, Lucía Teijerio, John Dines and my intern Vagia Tsiminaki. Some parts of this thesis would have not been possible without their expertise and their high-quality work. Also, thank you to Laurent, who "volunteered" to translate the abstract of this thesis to French.
Thank you to the Idiap Research Institute for providing such a beautiful working environment. Thank you to Nadine Rousseau, Sylvie Millius, and Edward Gregg for the excellent administrative support. Thank you to the Idiap System Team. Thank you to Olivier Bornet, who never failed to answer my technical questions, and to Hugo Penedones for very enjoyable discussions during commuting. A sincere thank you to Chantal Schneeberger, from LIDIAP, who was always kind to help, when needed at EPFL.
I feel lucky to have met amazing people during my years in Switzerland. Thank you to the "catalan clan" for making me feel at home, inside and outside the Catalan Center. Thank you to Pau Guri, Xavier Urbaneja, Josep Cuscó, Montserrat Ferrer, Ignasi Melià, Tona Fernández, Gerard Vinyes, Júlia López, Judit Fosses, Mariona López, Jordi Fernández, Maria Güell, Oriol Boada, Nuria Tenas, Esteban Bofill, Jorge Solana, Alioscia Hosch, and Alejandra Ramos. Thank you to Gerard Roca, Laura Paradell, and Olga Vinyals for oxygenating this team.
Thank you to another team of artists, geeks, hikers, runners, and dancers; always ready for action. Thank you to Ganga Garipelli, Margot Idrac, Gil Abrantes, Filipa Silva, Laura Capdevila, Leticia Carmo, Jessica Esteve, Oriol Fauria, Aline Favrat, Gerardo Chavez, Robert Leeb, Susana Limao, Jose Molina, Bernat Palou, Perrine Sigaud, Andrea Zakova, and Edgar Rangel. Also, thank you to my lunch partners, Ganga, Robert, Gerardo, Bernat, and more others, that accompanied my excursions at EPFL.
A big thank you to Miriam Guillamón, who appeared when I needed it he most. Also thank you to the Libertad Machordom for never saying 'No'.
Thank you to the people that helped me settle in Lausanne. Thank you to Javier Sanchez, Juan Perez, and Damien Mangialetto, with whom I discovered the city. A big thank you to Danielle Hulan for rocking our runs around the lake. Also, thank you to my former flatmates, Stephan Kenzelman, Julian Kellerhals, and Petri Alder, for the great times shared at Montolivet.
During my last year of PhD, I also benefited from visiting HP Labs, in Palo Alto, and Yahoo! Research, in Barcelona. Though the work there did not directly contributed to this thesis, these internships were an incredible opportunity to look outside my research topic and to meet very interesting people. Thank you to Bernardo Huberman for his friendship and to Alejandro Jaimes for valuable feedback.
My stay in San Francisco would have not been the same without some people. Thank you to Rubén Rodríguez for being the most generous person in Earth. He offered himself to drive me regularly to Palo Alto, even before meeting me. Also, thank you to Carlos Serrano and Hector Martín for being there again.
Back to Barcelona after five years, I felt almost like a stranger. Thank you to Mariona Querol, Rossano Schifanella, Joan Taberner, Asia Tximeleta, Luca Aiello, and Núria Verdiell who made the most of my time. Especially thank you to Mariona and Núria for teaching me that smiling

is a daily exercise that needs practice.
Some old friends have never stop following my adventures in the distance. Thank you to Núria Liras, Manuel Lozano, Pol Blasco, Jordi Fernández, Josep Puig, Víctor Balbastre, and Albert Mas. Also, a big thank you to Cristian Canton, without whose advice, I would have never started this journey.
The days previous to my defense were more intense than I would have predicted. Thank you to Judit for her support and valuable help. Thank you to Ganga for being the worst at giving feedback, and the best at giving advise. Thank you to Robert and Zahra Khalilli for their encouragement.
Finally, none of the work presented here could have been possible without the financial support of the Swiss National Science Foundation through the Interactive Multimodal Information Management (IM2) project.
To all of you, I can only say: thank you, thank you, thank you.

Lausanne, June 14th, 2013 Joan-Isac Biel

Abstract

The ubiquity of social media in our daily life, the intense user participation, and the explosion of multimedia content have generated an extraordinary interest from computer and social scientists to investigate the traces left by users to understand human behavior online. From this perspective, YouTube can be seen as the largest collection of audiovisual human behavioral data, among which conversational video blogs (vlogs) are one of the basic formats. Conversational vlogs have evolved from the initial "chat from your bedroom" format to a rich form of expression and communication that is expanding to innovative applications where a more natural and engaging way of reaching audiences is either necessary or might be beneficial. This video genre, available online in huge quantities, is a unique scenario for the study and characterization of complex human behavior in social media, that contrarily to social networks, text blogs, and microblogs, has remained unexplored so far.

The automatic behavioral understanding of conversational vlogs is a new domain for multimedia research. In short, the goal of our research is the understanding of the processes involved in this 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. While the study of vlogging contributes to user behavior research in social media, it also adds to a larger research agenda in social computing by analyzing behavioral data at scales not previously achievable in other scenarios. These type of analysis pose important challenges regarding the development and integration of methods for robust and tractable audiovisual processing.

In this thesis, we address the problem of mining user behavior inside conversational videos by addressing three main aspects. First, we integrate state-of-the art audio processing and computer vision techniques to analyze conversational social video. While the initial focus of the thesis is the nonverbal aspect of vlogger behavior, we also investigate the verbal content. Second, we study some of the interpersonal and social processes that link vlogger behavior and vlog consumption in social media platforms such as YouTube. In this context, we examine the phenomenon of social attention in vlogs, and we investigate the use of crowdsourcing as a scalable method to annotate large multimodal corpora with interpersonal perception impressions. Finally, we propose a computational framework to predict interpersonal impressions automatically using multimedia analysis, crowdsourced impressions, and machine learning techniques.

We anticipate that the work presented in this dissertation will motivate future work in social

and behavioral sciences, media analysis, natural language processing, and affective and social computing applied to the large-scale analysis of human interaction in social video.

keywords: social media, social computing, user behavior, online social video, video blogging, vlogging, interpersonal perception, first impressions, YouTube, social attention, personality, first impressions

Résumé

L'omniprésence des média sociaux dans notre vie de tous les jours, la participation intense des utilisateurs et l'explosion de contenus multimédia ont généré un intérêt extraordinaire chez les chercheurs en informatique et en sciences sociales dans le but d'investiguer les traces laissées par les utilisateurs pour comprendre le comportement humain en ligne. De ce point de vue, YouTube peut être considéré comme la plus grande collection de données multimédia à propos du comportement humain, parmi lesquels les vidéos blogs (ou vlogs) conversationnels en sont un des formats élémentaires. Les vlogs conversationnels ont évolué du format initial de "conversation depuis ta chambre à coucher" à une forme d'expression et de communication riche qui s'étend progressivement à des applications innovantes dans lasquelles une manière plus naturelle et engageante est soit nécessaire, soit avantageuse. Ce genre de vidéos, disponible en ligne dans des quantités gigantesques, offre un scénario unique pour l'étude et la caractérisation du complexe comportement humain dans les média sociaux, ce qui, au contraire des réseaux sociaux, des blogs textuels et des micro-blogs demeure inexploré à ce jour.

La compréhension automatique du comportement de vlogs conversationnels est un domaine nouveau dans la recherche multimédia. En résumé, le but de notre recherche est la compréhension des processus impliqués dans ce type de média sociaux, en ce basant non seulement sur le canal verbal - ce qui est dit - mais aussi sur le canal non-verbal - comment c'est dit. Le canal non-verbal comprend entre autres la prosodie, le regard, l'expression faciale, la posture, ou la gestuelle et a été étudié en profondeur dans le domaine de la communication non-verbale. En plus de contribuer à la caractérisation du comportement humain dans les média sociaux, l'étude des vlogs contribue aussi à un programme de recherche plus large dans le domaine de l'informatique sociale, en analysant des données comportementales à des échelles auparavant inatteignables dans d'autres scénarios. Ce type d'analyse pose des défis importants quant au développement et à l'intégration de méthodes de traitement audio-visuel robustes et faciles d'utilisation.

Dans cette thèse, nous abordons le problème de l'extraction du comportement de l'utilisateur de vidéos conversationnelles en abordant trois aspects principaux. Premièrement, nous intégrons techniques de pointe de traitement audio et de vision par ordinateur pour analyser des vidéos conversationnelles. Même si l'objectif initial de cette thèse est l'étude du comportement non-verbal des vlogueurs, nous investiguons aussi certaines caractéristiques provenant du comportement verbal. Deuxièmement, nous étudions plusieurs processus inter-personnels et sociaux qui relient le comportement des vlogueurs et la consommation de vlogs sur les

plates-formes de média sociaux telles que YouTube. Dans ce contexte, nous examinons le phénomène d'attention sociale dans les vlogs et nous étudions l'utilisation du 'crowdsourcing', c'est-à-dire l'externalisation ouverte, en tant que méthode d'annotation des corpus multimodales avec des avis interpersonnels à grande échelle.
Finalement, nous développons des modèles mathématiques qui peuvent prédire de manière automatique les impressions inter-personnelles en se basant sur des méthodes d'analyse multimédia, le crowdsourcing d'impressions et des techniques d'apprentissage par ordinateur. Nous anticipons que le travail présenté dans cette dissertation motivera des travaux futurs en sciences sociales et comportementales, en analyse de média, en traitement du langage naturel et en informatique affective et sociale, appliqués à l'analyse grande échelle d'interactions humaines dans des vidéos sociales.

Mots-clés : média sociaux, informatique sociale, vidéos sociales en ligne, blogs vidéos, vlogs, perception interpersonnelle, premières impressions, YouTube, attention sociale, personalité

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and objectives

Online video is more than YouTube, its ever-increasing rate of video uploads, and all the singers, dancers, actors, and musicians taking part on it. One particular format, conversational video, is constantly evolving and expanding to innovative applications (e.g. e-learning, question answering, online dating, or marketing testimonials) where a more natural and engaging way of reaching the audience is either necessary or might be beneficial. While all these emerging forms of video interaction are taking off, conversational video blogs (or in short, vlogs) have become a well established type of conversational social video, and one of the most prevalent formats among user-generated content in social media sites like YouTube [Burgess and Green, 2009].

The increasing ubiquity of social media in our daily life, the intense user participation, and the explosion of multimedia content available online have generated an extraordinary interest from computer and social scientists in understanding the ways in which people communicate in social media outlets such as Facebook, Twitter, or YouTube. Most efforts to study user behavior and interaction have largely focused on the study of social networks, blogging, and microblogging, primarily because of their popularity, but also because, compared to video, text and other metadata are easier to process at large-scale and there are plenty of analytical tools available [Kramer and Rodden, 2008]. However, despite the popularity of video online and the emergence of new forms of video communication and interaction, most online social video content remains unexplored from the behavioral perspective.

We posit that the automatic behavioral understanding of conversational vlogs is a new domain for multimedia research. In short, the goal of our research is the understanding of the processes involved in this 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]. While the study of vlogs contributes to the characterization of human behavior in social media, our research also contributes to a larger

human interaction modeling agenda in social computing [Pentland, 2008, Gatica-Perez, 2009]. In addition, the large-scale automatic analysis of vlogs poses important challenges regarding the development and integration of methods for robust and tractable audiovisual processing.

The overall goal of this thesis is the development of a framework to analyze social behavior in conversational vlogs through an interdisciplinary approach that combines social and personality psychology with computer science research. From the social psychology research perspective, our goal is to investigate some of the interpersonal perception processes that link vlogger behavior to vlog consumption in social media platforms such as YouTube. From the computer science perspective, we aim to exploit the potential of using audio processing and computer vision techniques in conjunction with crowdsourcing to build computational models of vlogger behavior.

The methodology used in this thesis as well as the language used during writing have been chosen to address both the social psychology and computer science communities. We believe that this thesis may be of great interest to social psychologists studying interpersonal perception and personality; social media researchers investigating user behavior; and social computing researchers studying human interaction in conversational scenarios.

1.2 Mining Multimodal Interaction in Social Video

The automatic analysis of multimodal interaction, and in particular of nonverbal behavior has been addressed by works in behavioral and ubiquitous computing, which have collected behavioral data using audio and visual sensors in multiple face-to-face scenarios. Grounded on classic research in communication theory [Knapp and Hall, 2005], these works have shown that nonverbal cues automatically extracted from "thin-slices" of behavior are useful to estimate functional roles [Zancanaro et al., 2006] and dominance in group meetings [Jayagopi et al., 2009], or the outcome of dyadic salary negotiations [Curhan and Pentland, 2007]. Though more difficult to extract automatically, the verbal behavioral aspect of face-to-face communication has also been investigated using manual transcriptions to predict the personality impressions of people in daily interactions [Mairesse et al., 2007].

Because of the amount of personal footage that people upload on YouTube, this site can can be seen as the largest collection of human behavioral "thin-slices". In this thesis, we consider the particular video genre that results from a single person talking most of the time in front of the webcam. We are interested in this configuration for two reasons. First, compared to other vlogging formats such as sketch-comedies, musical performances, or home scene footage, conversational vlogs are the ones that display the largest amount of conversational behavior. We deliberately refer to them as conversational instead of monologues because users tend to behave as if they were having a conversation with their audience through their webcam [Wesch, 2009]. Second, there is evidence that the single talking-head format accounts for a large proportion of user-generated video in YouTube, which indicates the availability of large-scale data [Burgess and Green, 2009].

A striking feature of conversational vlogging is the nature and variety of behavioral data that can be used to study and characterize both vloggers and their audiences. On the one hand, vloggers produce and share a diversity of (often highly personal) verbal content including opinions, desires, and personal narratives, and also a myriad of spontaneous audiovisual nonverbal cues though face, body, and voice. On the other hand, these videos are viewed by a potentially large audience that generates another multitude of behavioral traces, in this case from the watching crowd: views, comments, ratings, etc. the amount of behavioral data from the audience, and the large-scale amount of videos available, All taken together, vlogs offer a unique human interaction scenario that is relevant both to social media and social computing research.

In this thesis, we address the computational modeling of vlogger behavior in the framework of interpersonal perception: we investigate the automatic extraction of nonverbal and verbal behavioral cues that convey information from vloggers; the use of crowdsourcing as a scalable method to provide interpersonal impressions from vloggers; and the development of computational models that can predict interpersonal impressions automatically. To the best of our knowledge, our work is the first attempt to automatically analyze online social video from a computational behavioral perspective.

1.2.1 Summary of Contributions

The main contributions of this thesis are the following:

1. **We investigate the use of state-of-the-art multimodal techniques to extract behavioral cues from vlogs.** We show that current state-of-the-art techniques for audio processing, computer vision, and text processing are useful to automatically extract nonverbal and verbal behavioral cues from vlogs at large-scale (with some limitations). Though the the large-scale aspect is not fully exploited in our work compared to the sample sizes analyzed in other social media research, the amount of audiovisual data analyzed is larger by one order of magnitude than existing works analyzing face-to-face human interaction.

2. **We establish a number of links between automatic nonverbal cues and social attention.** We use correlation analysis to study the links between automatic nonverbal behavioral cues from vloggers and attention measures derived from YouTube metadata (e.g. view counts, number of comments, average ratings, etc), and we show that nonverbal behaviors from vloggers are strong determinants of social attention. To our knowledge, this is the first time such a connection is found in online social video.

3. **We propose the use of crowdsourcing and demonstrate its power to collect human impressions from conversational social video.** We design a new crowdsourcing experiment to collect interpersonal impressions from ordinary people during video-watching, and use reliability analysis to show that crowdsourced impression agreement compares

to annotations using traditional methods. Though the use of crowdsourcing has already been exploited for other human annotation tasks, to the best of our knowledge, our work constitutes the first attempt to crowdsource this type of personal and affective human impressions from conversational online video, and demonstrates that crowdsourcing is a systematic and scalable alternative to collect vlogger annotations from large-scale data.

4. **We examine the problem of building personality impressions from vlogs.** We use a cue utilization analysis to study how personality impressions mediate the vlog watching experience connecting vlogger behavior to social attention. First, we show that crowdsourced impressions are correlated to social attention in YouTube. Second, we show that certain nonverbal and verbal cues extracted automatically from audio and video are correlated to interpersonal impressions from vloggers, in ways that concur with previous social psychology and social media research.

5. **We propose a computational framework to predict vlogger impressions.** We approach the problem of automatically predicting vlogger impressions using automatic multimodal nonverbal and verbal cues, crowdsourced annotations, and machine learning techniques. First, we address the task of prediction personality impressions and show that nonverbal and verbal content models are successful at predicting Big-Five trait impressions, with varying results across modalities and traits. Second, we investigate the prediction of vlogger impressions beyond personality by exploring an alternative, multifaceted representation that is data-driven and is extracted using probabilistic topic models.

1.2.2 Thesis Outline

The rest of the thesis is organized in five chapters:

In Chapter 2, we review the literature on five different areas from social media, multimedia, and behavioral computing research that relate to our work.

In Chapter 3, we address the problem of automatically extracting nonverbal behavior from vloggers. This chapter discusses the use of state-of-the-art audio processing and computer vision techniques to process vlogs automatically and to study the links that exist between vlogger behavior and social attention. This chapter also introduces our data and provides insights about the type of editing elements involved in conversational vlogging.

In Chapter 4, we address the problem of personally impressions in vlogging. In this chapter we investigate several characterizations of vlogger behavior based on nonverbal cues from speaking activity, prosody, head and body activity and facial expressions of emotion, and verbal content features extracted from manual and automatic transcriptions of the vloggers' speech. We study correlates between different features and personality impressions, and we address the task of automatic personality impression prediction.

In Chapter 5, we present our work on crowdsourcing interpersonal impressions from vloggers. The first part of the chapter presents the annotation experiment designed to crowdsource vlogger impressions and the reliability analysis performed on the collected annotations. In the rest of the chapter, we investigate interplays among multifaceted impressions, and propose the use of an unsupervised method (probabilistic topic models) to identify prototypical impressions of vloggers beyond personality alone. We also address the problem of automatically predicting prototypical impressions using nonverbal features, verbal features from manual transcripts and verbal features extracted from YouTube comments.

In Chapter 6, we conclude by reviewing the main contributions and findings of this thesis, discuss some of the limitations of our work, and propose future research directions.

1.2.3 Publications

Most of the work presented here was published in the following peer-review journal articles and conference papers (listed in inverse chronological order):

Journal Articles

J.-I. Biel, D. Gatica-Perez. The Youtube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs. *IEEE Transactions on Multimedia*, 15(1): 41-55, 2013.

J.-I. Biel, D. Gatica-Perez. VlogSense: Conversational behavior and social attention in YouTube. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP), Special Issue on Social Media*, 7(3):33:1-33:21, 2011.

Book Chapter

J.-I. Biel, D. Gatica-Perez. Call me Guru: user categories and large-scale behavior in YouTube. *Social Media Modeling and Computing*, 2:167-188, 2011.

Conference papers

J.-I. Biel, V. Tsiminaki, J. Dines, D. Gatica-Perez. Hi YouTube! Personality Impressions and Verbal Content in Social Video. *(submitted for conference publication)*, 2013.

J.-I. Biel, L. Teijeiro-Mosquera, D. Gatica-Perez. FaceTube: predicting personality from facial expressions of emotion in online conversational video. *Proceedings of the 14th ACM international conference on Multimodal Interaction (ICMI)*, 2012.

J.-I. Biel, D. Gatica-Perez. The Good, the Bad, and the Angry: Analyzing Crowdsourced Impressions of Vloggers. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.

J.-I. Biel, O. Aran, D. Gatica-Perez. You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2011.

J.-I. Biel, D. Gatica-Perez. Vlogcast yourself: Nonverbal behavior and attention in social media. *Proceedings of the 13th ACM International Conference on Multimodal Interfaces (ICMI-MLMI)*, 2010.

J.-I. Biel, D. Gatica-Perez. Voices of vlogging. *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

J.-I. Biel, D. Gatica-Perez. Wearing a YouTube hat: directors, comedians, gurus, and user aggregated behavior. *Proceedings of the ACM International Conference on Multimedia (MM)*, 2009.

Others

J.-I. Biel. Please, subscribe to me! Analyzing the structure and dynamics of the YouTube network. Unpublished. EPFL Dynamical Networks Course project. Autumn, 2009 (available at http://www.idiap.ch/~jibiel/pubs/Biel09YouTubeSubscr.pdf).

2 Related work

In this chapter, we review the literature on five different areas in social media, multimedia, and behavioral computing research that relate to our work on mining conversational social video.

First, our research contributes to the investigation of the different facets of YouTube as a sociotechnical system. The literature review in Section 2.1 provides a broad picture of the diverse research perspectives, the focus being put on videos and users, and the specificity and scale of the analysis used to investigate YouTube. To the best of our knowledge, our work is the first attempt to use automatic content analysis of social video to characterize the nonverbal and verbal behavior of users, as opposed to other attempts focused on users that have analyzed the textual metadata and social network links.

In Section 2.2, we refer to the phenomenon of attention in social video. In particular, we review the different approaches to investigate patterns of popularity in YouTube, as well as research investigating the impact of attention in the production of user-generated video. In this context, we discuss our approach to the study of attention in vlogging, focusing on the social nature of attention and the conversational aspect of vlogging, and we refer to work investigating related interpersonal perception concepts in face to face interactions.

In Section 2.3, we review the growing body of literature in social media focused on characterizing different aspects of user personality, attractiveness, and mood states, by either automatically mining their manifestations or from the perspective of interpersonal perception research. The review makes evident that social video is an understudied format in social media research, and highlights how our work extends the literature by examining the audiovisual modalities of social media user behavior.

As discussed in Section 2.4, our research also contributes to recent efforts that explored the feasibility of crowdsourcing the annotation of multimodal corpora or that have used Mechanical Turk to carry out human studies. Compared to the related literature, our work constitutes a first attempt to crowdsource the annotation of interpersonal impressions from

video watching.

Finally, in Section 2.5, we review behavioral and ubiquitous computing works that have addressed the automatic analysis of human behavior in face to face conversational scenarios to model social and personal constructs. In contrast to all these works, we analyze a new form of human communication that can be analyzed at a larger scale and that includes a huge diversity of human behaviors.

2.1 Analyzing Online Social Video

YouTube admits multiple conceptualizations of what it is and what it is for, and each of these perspectives have motivated different approaches when taking YouTube as a research subject [Burgess and Green, 2009]. In this section, we review different ways of posing questions around YouTube in different areas of network systems, social media, multimedia, new media, and ethnography.

A number of research works in computer science have treated YouTube as a video distribution system to investigate the impact of user-generated content in underlying video-on-demand architectures. These works have analyzed millions of videos from YouTube and have studied how the properties of user generated video and the macroscopic characteristics of the social network [Cha et al., 2007, Cheng et al., 2008], and have also investigated the problem of YouTube video distribution in local networks [Gill et al., 2007, Zink et al., 2008]. This view of YouTube has also motivated an interest in understanding the phenomenon of video popularity [Szabo and Huberman, 2010], that we review in more detail in Section 2.2.

Research has also investigated YouTube as the paradigm of participatory culture around video. With videos and content at the center of analysis, new media and communication research has investigated the type, the properties, and the editing constituents of YouTube videos, in order to gain understanding on user-generated media production [Landry and Guzdial, 2008], and to investigate the shift from traditional media to user-generated content consumption [Halvey and Keane, 2007, Burgess and Green, 2009]. These works have manually coded thousands of videos to categorize the types and properties of YouTube content, while others have studied smaller samples of videos to manually analyze the discursive aspect of videos, their visual content, and the reactions of people to understand video production and consumption [Molyneaux et al., 2008, O'Donnell et al., 2008].

Other research works focusing on users have analyzed YouTube metadata generated by thousands or millions of users to understand long-term behavior [Kruitbosch and Nack, 2008, Biel and Gatica-Perez, 2009], and have also investigated the nature of social interactions through social network analysis based on friendship and subscription links [Mislove et al., 2007, Biel, 2009], and the use of video responses [Benevenuto et al., 2009]. Some of these prior findings have been recently backed up by a detailed analysis of the full-scale YouTube subscription, comment, and video graphs [Wattenhofer et al., 2012].

YouTube has also been subject of research in long-term ethnographic studies that have observed the offline behavior of YouTube teenage users [Lange, 2007a], and have also get involved in creating videos, leaving comments, and "friending" YouTube vloggers [Wesch, 2009]. These works provided key insights on why people use YouTube in they daily life [Lange, 2007a], what is the influence of feedback, criticism and hate behaviors [Lange, 2007b], and what forms of awareness exist in relation to the self and the others [Wesch, 2009]. The problem of creating an online identity through video was also investigated from the perspective of new media studies [Griffith, 2007].

Understood as a massive video repository, the scale and diversity of YouTube content poses many interesting challenges to solve several multimedia-centered tasks. From this perspective, research has addressed the detection of near-duplicate video using audiovisual analysis [Cherubini et al., 2009, Oliveira et al., 2010], the detection of extremist and hate content [Sureka et al., 2010], or the automatic modeling of the video content topic and ideological perspective [Lin and Hauptmann, 2008]. The automatic analysis of videos has been also used to learn to classify videos at large-scale [Wang et al., 2010b], and to investigate methods for automatic event summarization [Hong et al., 2009], classification [Ni et al., 2011, Xie et al., 2011], and tracking [Xie et al., 2011].

The scale and diversity of YouTube have also evidenced the need to advance traditional audiovisual analysis technologies. Along this line, recent works have used YouTube as a testbed to evaluate face tracking [Kim et al., 2008], speech/non-speech segmentation [Misra, 2012], and automatic speech recognition technologies [Hinton et al., 2012].

Our work contributes to the understanding of YouTube as a collection of communities where individuals express and communicate through videos. Compared to the study of general samples of YouTube videos [Cha et al., 2007, Cheng et al., 2008, Wattenhofer et al., 2012] and users [Mislove et al., 2007, Benevenuto et al., 2009, Wattenhofer et al., 2012], we focus on vloggers, who explicitly show themselves talking in front of the camera. However, despite the high popularity of this video genre in YouTube [Burgess and Green, 2009], very few of the works cited here address this format in their studies. Some of the works in new media, communication, and ethnography have conceptualized vlogs in a broader sense than the pure conversational one, but in practice they consistently reported that the major part of the content as featuring a single participant talking to the camera [Lange, 2007a, Griffith, 2007, Molyneaux et al., 2008, Wesch, 2009]. In addition, our work differs from other works that have used the term *videoblog* or *vlog* in the sense of “a log of videos”, instead of a conversational vlog [Zhang et al., 2009].

Compared to other works focusing on users, our research goes beyond the behavioral interpretation of metadata [Kruitbosch and Nack, 2008, Biel and Gatica-Perez, 2009] and relationship links [Mislove et al., 2007, Biel, 2009, Benevenuto et al., 2009], to focus on the actual behavior displayed by users in their videos. In this sense, our work relates to multimedia attempts to automatically analyzed video [Oliveira et al., 2010, Sureka et al., 2010] with a specific focus on

the human behavioral aspect of the analysis. Though the large scale aspect of YouTube [Wang et al., 2010b] is not fully exploited in our research, our work shows how state-of-the art technologies can be used for the computational understanding of vlogs. In addition, the vlogging scenario itself represents a test bed for the development of audiovisual technologies that specifically attempt to measure human behavior.

2.2 Video Popularity and Social Attention

The sheer amount of content in video sharing sites such as YouTube and the limited amount of time that people have to watch videos have awaken an interest in understanding why and how certain videos become popular while others are rarely watched. Beyond the clear economic benefit for content creators and providers on determining the virality of content ahead of time, understanding and detecting popularity has potential impact on a wide range of systems and applications, spanning from delivery networks to recommendation and discovery engines [Broxton et al., 2010].

Early works investigating the popularity of YouTube videos focused on the statistical analysis of different characteristics of videos, the social network properties of YouTube [Cha et al., 2007, Cheng et al., 2008], and the coarse evolution of video view counts [Cha et al., 2007, Szabo and Huberman, 2010], or have alternatively measured the network traffic generated by popular videos [Zink et al., 2008]. In addition, research has investigated the prediction of video popularity based on the initial number of views [Szabo and Huberman, 2010].

A more comprehensive picture of video popularity was provided by research dissecting the fine-grained evolution of video views [Broxton et al., 2010]. By looking at the source of these views, this work unveiled the complex effects of video sharing (inside and outside YouTube), and video search on the growth and decay of views for non-popular, long-term popular, and viral videos. In addition, recent research has also investigated the geographical reach of videos around the world based on the daily evolution of views during the videos life-time [Brodersen et al., 2012].

The phenomenon of video popularity responds to a new economy, where attention is scarce and becomes the valuable good that people seek [Goldhaber, 1998]. In a recent study, Huberman et al. [2009] investigated how the pursuit for attention impacts the productivity of content creators in social media. In particular, they showed that the productivity of users uploading videos to YouTube strongly depends on the amount of attention received in previous uploads, measured by the log number of view counts. In addition, they showed that a lack of attention leads to a decrease in the number of videos uploaded and a consequent drop in productivity.

In our work, we address the problem of attention in vlogging from a perspective shared between the new economy of attention and a more traditional view of this phenomenon in the study of human interaction. In this context, we use the term "social attention" to refer to the social aspect of attention. On the other hand, we exploit the metadata of videos (view counts,

number of comments, average ratings, etc) to measure the intensity of the attention received by vloggers online, as the aforementioned works studying popularity in YouTube. On the other hand, by focusing on the human communication aspect of vlogs, we aim to examine the attention phenomenon in vlogs grounded on interpersonal perception theory [Ambady and Rosenthal, 1992]. Along this line, our work relates to existing literature investigating personal, social, and behavioral aspects related to achieving attention from other people or reacting to it. This includes prior research in social psychology linking the popularity of extraverts to their desire to captivate the attention of others, their expressive behaviors, verbal humor, and fashionable dress code [Ambady and Rosenthal, 1992, Borkenau and Liebler, 1992, Ashton et al., 2002, Scherer, 1979] in addition to works that relate these cues to emotional expressivity and social dominance [Butler et al., 2003, Dovidio and Ellyson, 1982b].

2.3 Characterizing Social Media Users

The automatic analysis of vlogs, as addressed in our work, extends within the audiovisual domain the growing body of studies in social media focused on blogs, microblogs, and online social networks that seek to provide insights on user behavior. More specifically, our research relates to works that have investigated different aspects of users' personality, attractiveness, and mood, either from the perspective of interpersonal impressions or by directly mining their explicit verbal expression.

2.3.1 Personality and Attractiveness Impressions in Social Media

Works investigating the processes of interpersonal perception in social media have mainly focused on the use of personality traits as broad descriptors of user characteristics. This initial approach has been partly motivated by the wide acceptance of the Big-Five model as a way to organize people's traits, the existence of standard measures of personality, as well as a tradition of personality research in social psychology.

Early personality research in social media suggested that people tend to use social media to express and communicate their personality Gosling et al. [2007] and reported that a high percentage of social network users explicitly manifest their desire of representing their personality traits [Counts and Stecher, 2009]. This research has used brief measures of personality [Gosling et al., 2003] to measure both self-reported personality and personality impressions in order to investigate how people present themselves in social networks; how they convey personal information when creating their user profiles; and how are they seen by others based on their behavior.

In one of the first studies of personality in social media, [Gosling et al., 2007] showed that personality impressions inferred from Facebook user profiles achieved significant agreement among observers, which suggests that user profiles convey useful personality information. In addition, they showed that these impressions were generally accurate with the profile owners

self-reported personality, regardless of whether this personality is idealized or not [Gosling et al., 2007]. In this thesis, we extend the study of personality impression agreement to social video, but we do not address the study of impression accuracy. Instead, we focus on studying the personality impression correlates that can be found in social media content. In this line, previous research, has investigated user profiles to identify what elements from text and pictures are associated to more or less accurate personality impressions and also consistently showed females to elicit higher impression agreement [Evans et al., 2008, Steele Jr et al., 2009]. As a general note, none of these works used involved any automatic analysis of multimedia content.

The study of personality impressions has also been addressed in blogs [Li and Chignell, 2010]. In particular, the work was focused on linking bloggers writing style with personally impressions, and showed that personality impressions inferred from reading blogs also achieved readers' agreement. In addition, this work showed that different types of blogging elicit different levels of impression accuracy.

Prior interpersonal perception research has also investigated some aspects of attractiveness impressions in social media, as we address in this thesis. In particular, research in online dating sites has investigated what elements of user profiles are associated with different facets of attractiveness [Fiore et al., 2008], and has also investigated mate preference based on user attributes such as income, education, physique, and physical attractiveness, as well as information on the users' religion, political inclination, etc. [Hitsch et al., 2005]. The problem has also been addressed in social networks, where works have focused on studying the attractiveness impression information conveyed by Facebook user profiles. For example, works have investigated to what extent the attractiveness of users is influenced by both the activity generated by friends in the users' wall and the attractiveness of these friends [Walther et al., 2008], or by the number of friends alone [Tong et al., 2008]. Other research has shown that people prefer attractive users when initiating online relationships with zero history [Wang et al., 2010a].

2.3.2 Mining Social Media User Personality and Mood

Different from interpersonal impressions, research in social media has also characterized user personality and mood by directly mining verbal expression in blogs, microblogs, and social networks.

Previous research has examined the links between word usage and self-reported personality in blogs as opposed to personality impressions, as in the previous section. In particular, Gill et al. [2009b] automatically analyzed a sample of 2,400 blogs and 12M words to study the relation between people self-reported personality and writing styles, and the relation between personality and motivations for blogging. The same problem was addressed by Yarkoni [2010] in a larger sample of 700 bloggers and 80M words, who provided a more comprehensive study of the actual associations between word usage and bloggers' personality. Overall, these works are

great examples of how large-scale social media data can be used to back up earlier results linking individual differences on personality and linguistic styles in social psychology [Pennebaker and King, 1999].

The above works have been followed by several attempts to automatically predict the personality of bloggers [Oberlander and Nowson, 2006, Nowson and Oberlander, 2007]. First, Oberlander and Nowson [2006] investigated the use of an n-gram model to automatically classify the self-reported personality of 70 bloggers, achieving accuracies between 75% and 84% on a balanced, binary classification task. However, when the same approach was used to classify a larger sample of 1769 blogs, classification accuracies decreased down to 52% and 59% respectively [Nowson and Oberlander, 2007]. The authors highlighted the difficulty of classifying personality using noisy, large-scale data from blogs and personality scores with relatively low reliability (whereas the questionnaire used in the first experiment included 41-items, the second questionnaire included only 5 items). Independently of whether the authors were interested on predicting the self-reported personality of bloggers rather than the impressions that readers make about them, these works show the potential of using automatic techniques for the prediction of personality in large-scale social media data.

Several works have also investigated mood expression in blogging, using large collections of blog posts and self-reported mood labels obtained from the bloggers themselves as they were blogging [Mishne, 2005]. These works have proposed different approaches to estimate blogger mood using automate text analysis, with promising results [Keshtkar and Inkpen, 2009, Leshed and Kaye, 2006, Nguyen et al., 2010].

In addition, the large amount and the spontaneity of mood expression in social media has generated an interest on harvesting and using social media data as coarse social sensors of mood at community and society scales. In the case of blog data, research has explored the temporal aspect of expression to relate the semantics of daily activities and interactions with people happiness [Mihalcea and Liu, 2006]. Other related works include exploiting the aggregates of Facebook status updates to obtain a gross happiness measure [Kramer, 2010], mining mood expression in blogs over time to identify seasonal trends [Balog and de Rijke, 2006], or estimating mood from tweets in relation with public social, political, cultural, and economic events [Bollen et al., 2010].

Summing up, related research has been prolific at the understanding of user profile content, photos, and text as drivers for self-presentation, personality and attractiveness impressions, as well as mood expression. However, to the best of our knowledge, our work is the first attempt to understand interpersonal perception in online social video with respect to the nonverbal and verbal behavior displayed by vloggers. This is relevant because the nonverbal channel conveys information that is often unconscious and is more difficult to control than profile information or language style [Gosling et al., 2007]. In addition, not much is known about the spontaneity and verbal discourse of vloggers compared to blog data and the feasibility of automatic analysis vloggers' verbal content. In addition, our work goes beyond the individual

focus of most social media literature that has investigated users' traits and states individually, by examining the interplay between different facets of vloggers.

2.4 Crowdsourcing Human Impressions

Crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) make it possible today to use human workforce online and solve highly complex tasks thanks to the collective knowledge and participation gathered from the small contributions of individual workers. The low cost of crowdsourcing labor has led to an unprecedented number of research efforts to exploit human generated data for tasks such as manual classification/categorization, data verification and correction, speech collection, image annotation, opinion mining, brainstorming, etc.

Along the line of recent literature showing the utility of micro-tasks markets for conducting human behavioral studies [Kittur et al., 2008], our work contributes to recent efforts that have explored the feasibility of crowdsourcing the annotation of multimodal corpora based on human observations [Soleymani and Larson, 2010, Brew et al., 2010]. Clearly, these tasks differ from other crowdsourced annotations in that there is no clear cut ground truth data.

The study of interpersonal perception requires the collection of human impressions from third-party observers. In some works, these observations can be collected among participants during the studies, specially on face-to-face interactions [Ambady et al., 1995]. However for the study of personality impressions using multimodal corpora, a common approach consists on gathering people in laboratory settings and ask them to listen/watch the data, and to judge the personality of the people on them [Mairesse et al., 2007, Mohammadi et al., 2010]. This approach, which is used in general to annotate other personal and social constructs in data corpora [Jayagopi et al., 2009], can become expensive in terms of time and money, specially if one aims to annotate large-scale data.

Previous research has used crowdsourcing for multimedia content annotation. Soleymani and Larson [2010] explored the use of MTurk to annotate the affective response of people to a set of 126 videos. They concluded that crowdsourcing was a valuable technique to collect affective annotations and provided a short guide of best practices to use MTurk for that purpose In the context of social media, Brew et al. [2010] crowdsourced the annotation required to train a machine learning system to score and track news feeds' sentiments, and investigated how to manage the effort of annotators to maximize the coverage and the agreement achieved in the annotations. Similarlty, other research has crowdsourced the annotation of microblogs' sentiment polarity [Diakopoulos and Shamma, 2010],

Related to the collection of personality data from crowdsourcing, Buhrmester et al. [2011] explored the collection of data for psychology studies by administering a series of self-reported personality questionnaires to MTurk workers, and found that the quality of data met the psychometric standards associated with published research. In addition, the results suggested that MTurk participants are at least as diverse and more representative of non-college popula-

tions than those of typical Internet and traditional social psychology samples. This is indeed an interesting feature of MTurk for our collection of personality impressions, because our purpose is to collect observations made by ordinary people (as opposed to trained annotators).

Compared to the literature, our work constitutes a first attempt to crowdsource interpersonal impressions from online social video in a framework that is suitable for the annotation of large-scale data, and that resembles a diverse online community who watches appealing content and disregards uninteresting one.

2.5 Automatically Modeling Human Behavior from Audio and Video

Our research adds to a series of works in behavioral and ubiquitous computing that use automatic nonverbal behavioral cues from audio and video in several face-to-face communication scenarios, and that are inspired by classic social psychology works on social perception and nonverbal behavior [Scherer, 1979, Dovidio and Ellyson, 1982a, Iizuka, 1992, Ambady and Rosenthal, 1992, Knapp and Hall, 2005].

These works have shown that automatically extracted nonverbal cues 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]. Prior research has shown that the automatic analysis from "thin-slices" of nonverbal cues from audio and video leads to good estimators of functional roles [Zancanaro et al., 2006] and dominance in group meetings [Jayagopi et al., 2009], and to reliable predictors of the outcome of salary negotiations [Curhan and Pentland, 2007].

Our work relates to previous attempts to automatically analyze audio and video for personality prediction. In this context, Mairesse et al. [2007] investigated the classification, regression and ranking of both self-reported personality and personality impressions using 96 audio segments captured in daily interactions. Using prosodic cues, their best results on the regression tasks were achieved for Emotion Stability and Extraversion with R^2 = 18% and R^2 = 8% respectively. Interestingly, their regression model could not predict self-reported personality.

In a related study, Mohammadi et al. [2010] focused on the use of prosodic cues to automatically classify Big-Five personality impressions obtained on French professional radio broadcasts (7h of data). The obtained accuracy ranged from 65% to 80% depending on the traits, and the authors reported the Extraversion trait as the easiest one to predict.

Lepri et al. [2009, 2010] also investigated the automatic prediction of self-reported extraversion and locus of control in small group meetings. They explored the use of audio cues (speech activity and prosody) and visual cues (energy from head, hands, and body) extracted on one-minute slices in a total of 6h of data. In their regression task, they obtained up to R^2 = 22% for Extraversion [Lepri et al., 2009]. They also found that using the gaze of others as a cue could help to improve the prediction [Lepri et al., 2010]. Previous work also used automatic

analysis of group meetings to computationally model dominance, a trait typically related to Extraversion [Hung et al., 2008, Jayagopi et al., 2009]. Jayagopi et al. [2009] investigated audio-visual cues for dominance estimation using both unsupervised and supervised models and found that speaking time led to superior performance. Also in group meetings, a high ratio between looking while speaking and looking while listening [Dovidio and Ellyson, 1982a] was found to determine dominance [Hung et al., 2008].

Recent work also addressed the study of personality from automatic audiovisual analysis in a somewhat similar monologue setting to vlogging [Batrinca et al., 2011]. However, although the audiovisual techniques used for analyzing such content may be similar to the ones used in our work, the data used and the task addressed are different in nature. First, as opposed to the recordings made in a laboratory [Batrinca et al., 2011], our data consists of spontaneous, self-recorded vlogs that are made to be shared publicly online. Second, we address the problem of personality impressions and not self-reported personality.

Regarding the verbal aspect of face-to-face interactions, some works have also investigated language usage in transcriptions of recorded daily interactions to study the links between everyday expression and both self-reported and impressions of personality [Mehl et al., 2006]. In addition, they have also added text to acoustic features for the task of automatically predicting personality, and obtained better results than using prosodic features alone for Extraversion (R^2 = 23%), Emotional Stability (R^2 = 9%) and conscientiousness (R^2 = 18%) [Mairesse et al., 2007].

Finally, our work relates to social psychology literature that has documented the face as an important source of information in interpersonal impressions [Knapp and Hall, 2005, Knutson, 1996, Hall et al., 2011]. These works have shown that people rely heavily in facial cues to make interpersonal judgements because there is a general belief that faces provide valuable information about a person's character or personality [Knapp and Hall, 2005]. In addition, our research relates to works that evidenced that facial expressions of emotion provide information other that emotional states, influencing interpersonal impressions such as personality judgments, and that specific affective cues are correlated with the possession of various personality traits [Knutson, 1996, Hall et al., 2011].

In contrast to all the above works, we analyze a new form of human communication, that resembles face-to-face interaction and that is potentially interpreted by a huge audience. Videoblogging has unique features, including a 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 having a conversation with their audience through their webcam [Wesch, 2009]. In addition, as opposed to the controlled settings and the high-quality sensors used in most current face-to-face behavioral research [Jayagopi et al., 2009, Zancanaro et al., 2006], vlogs result from widely varying processes 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.

2.6 Conclusions

Research in social media has investigated many different aspects of video-generated and user behavior. The literature review throughout this chapter has shown how the many different perspectives, analytical methods, and the specificity and scale of data used contributed to the understanding of social video and to the computational modeling of users personal traits and states. However, the complexity of human behavior and the variety of social media formats leave plenty of open problems. In the rest of this thesis, we present our approach to leverage behavioral and ubiquitous computing techniques to mine conversational social video. To the best of our knowledge, this is the first attempt to use automatic content analysis to characterize social video users based on the nonverbal and verbal behavior displayed in their videos.

3 Nonverbal Behavior and Social Media Attention

3.1 Introduction

In this chapter, we investigate the use of state-of-art audio processing and computer vision techniques to mine vlogger behavior. We focus on the nonverbal aspect of behavior, which has been investigated in previous literature in other conversational settings, and that, compared to verbal content, has potential for being robust to the variety of topics and language used in vlogs. To the best of our knowledge, our work is the first one that characterizes the spontaneous conversational nonverbal behavior in online social video.

The large-scale analysis of vlogs poses important challenges regarding the development and integration of methods for robust and tractable audiovisual processing. The first goal of this chapter is to integrate automatic feature extraction methods used in the social computing literature to characterize audio, visual, and multimodal activity cues in order to measure the variety of vlogger behaviors in a sample of YouTube vlogs. We leverage the output of the automatic processing pipeline to advocate for the usefulness of this characterization through our experiments, which include an analysis of the type of video editing and non-conversational content used in vlogs.

The second goal of this chapter is to investigate the importance of vlogger behavior with respect to the attention that vloggers receive in YouTube. Previous interpersonal perception research has shown that certain personal, social, and behavioral aspects are associated to people receiving attention from other people or reacting to it [Ambady and Rosenthal, 1992]. While this phenomenon has been documented in face-to-face interactions, we hypothesize that similar mechanisms may occur in vlogging. In our study, we investigate the behavioral correlates to an aggregated estimate of attention computed from the views counts of vlogs in YouTube. View counts are useful to measure the strength of attention in social video because they result from the aggregated behavior of millions of people watching videos online, and are used in related research investigating the popularity of online video [Cha et al., 2007, Cheng et al., 2008, Broxton et al., 2010]. In this context, we use the term *social attention* instead of popularity to refer to the social aspect of attention, but also to differentiate from attempts to

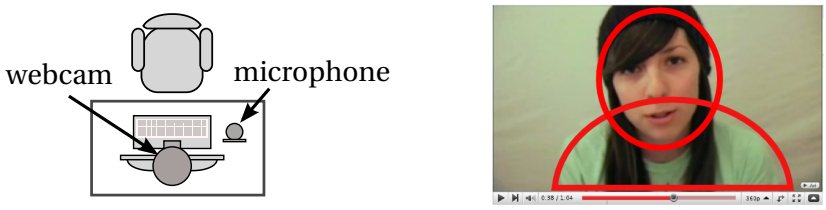

Figure 3.1: The basic vlog setup: a camera, a microphone (left), and a talking-head (right).

predict the popularity of YouTube videos [Cheng et al., 2008, Szabo and Huberman, 2010], as this is not the goal of our work.

The main contributions of the chapter are summarized as follows:

- We use state-of-the art audio processing and computer vision techniques to characterize vloggers' nonverbal behavior. We propose a principled way to segment and identify the conversational parts of vlogs, and to extract automatic audio, visual, and multimodal nonverbal cues motivated by social psychology, and applicable at large-scale.
- We use correlation analysis to establish the connection between nonverbal behavioral cues and aggregated estimate of attention computed from the YouTube views counts of vlogs. We show that specific patterns of vlogger behavior are related to the social attention that vlogs receive from their audience. To the best of our knowledge this is the first time that this result is reported in online social video.
- We analyze the output of the automatic processing methods to study the video creation and edition elements involved in vlogs when compared to other types of online social video, and show that the content found in vlogs is driven by a main conversational intent.

The rest of chapter is organized as follows. In Section 3.2, we present our data collection. In Section 3.3, we introduce our approach to the automatic processing of vlogs and extraction of audiovisual nonverbal cues. In Section 3.4, we present our experiments and discuss the results. We conclude and discuss some limitations of our work in Section 3.5. This chapter was published in [Biel and Gatica-Perez, 2011].

3.2 The YouTube dataset

Our collection of YouTube vlogs consists of 2,269 videos featuring one single person talking in front of the camera, with no restriction in terms of the topics addressed by vloggers, which is diverse, and that includes personal matters, movie, books, and product reviews, political debate, etc. Visually, these vlogs are simply identified by a talking-head occupying most of the screen, as illustrated in Figure 3.1). We choose this configuration for two main reasons.

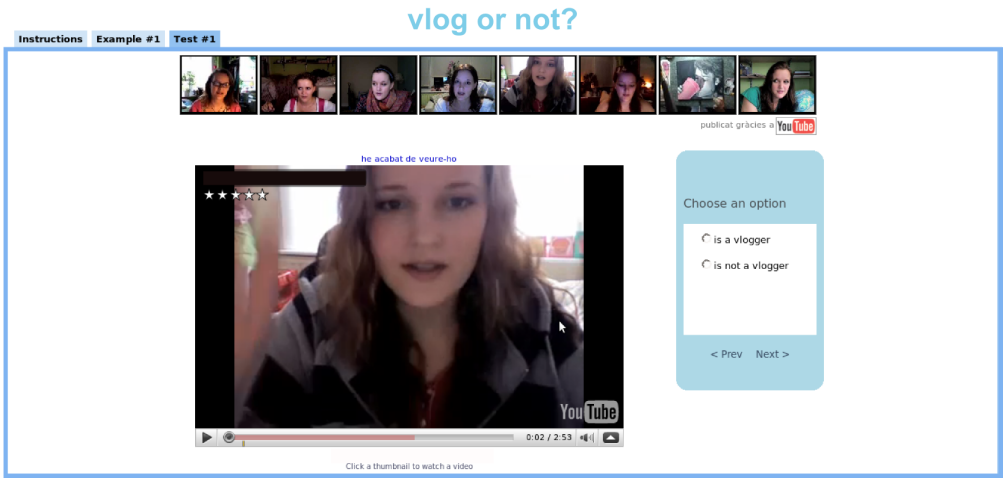

Figure 3.2: A view of the web application designed to annotate YouTube vlog collections. On the top, the video bar showing the last uploaded videos form the user. On the bottom (left), the vlog under inspection. On the bottom (right) the annotation form.

First, compared to other types of vlogging such as sketch-comedies, musical performances, or home scene footage, the single talking-head format is the one whose content is mostly conversational. 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 suggests the availability of large-scale data.

3.2.1 Data collection

We used a semi-automatic process to gather conversational vlogs from YouTube. First, we used the YouTube API to automatically query vlogs using the keywords “vlog”, “vlogging”, and “vlogger”. The variety of content and topics in vlogs is associated to a multitude of other keywords that can be used to search for vlogs but that also overlap with many other types of online video. A manual inspection of 300 randomly sampled search results showed that, using the above keywords, 25% of the videos retrieved corresponded to the talking-head setting. 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 use conversational vlogs regularly as a core practice, which results in vlog collections of a substantial size, other users vlog as a side activity.

As a follow up, we extracted a list of 878 different *usernames* from videos retrieved using the aforementioned keywords in November 2009 and built a web application (see Figure 3.2) to watch and annotate their 8 most recent videos. In particular, we asked 10 annotators to browse the videos using the progress bar, instead of watching videos completely, and to answer a few questions about the content. The questions aimed to identify the presence of the talking-head setting, the number of unique persons featured in the vlog, and the main conversational aspect (as opposed, for example, to the vlogger playing music or singing). Annotators were untrained volunteers whose only requirement was to be familiar with YouTube as a video viewer.

A total of 6,396 videos were annotated with one annotation per video (some users had less than 8 videos). Typically, each volunteer spent one hour to annotate the videos corresponding to 25 users. Out of these annotations, we selected videos featuring one person talking in front of webcam.

3.2.2 Dataset description

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 3.3 shows three aspects of our collection: the distribution of videos per vlogger, the duration of videos, and the view counts received by videos on YouTube at the time of collection. The distribution of conversational vlogs per vlogger in the dataset (Figure 3.3, left) is close to the average of five vlogs ($mean$ = 4.8, $median$ = 5). These vlogs have typical durations between 1 and 6min (70% of the videos appear in this interval), with a median duration of 3min 15s (Figure 3.3, center). Only 2.4% of the videos are longer than 10min, a time limitation that in 2009, could be only exceeded by YouTube partners (users who participate in the advertising program of YouTube). This feature 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 used in behavioral computing works [Gatica-Perez, 2009].

Our sample has also some desirable characteristics from the point of view of the study of attention. As shown in Figure 3.3 (right) the number of views per vlog varies linearly in the log-log scale, which is common in online video data and social network data in general, were a small percentage of the vloggers are very popular and a lot of them are ordinary vloggers that receive very low numbers of views. In our sample set, the distribution of views is skewed towards a small number of views ($median$ = 231, $mean$ = 20030). This relatively low value may be the result of collecting the most recent upload of each vlogger. In addition, we found that some of the most popular vloggers in our sample are also featured as popular users in YouTube, which indicates that the sample covers both extremes of the actual views distribution

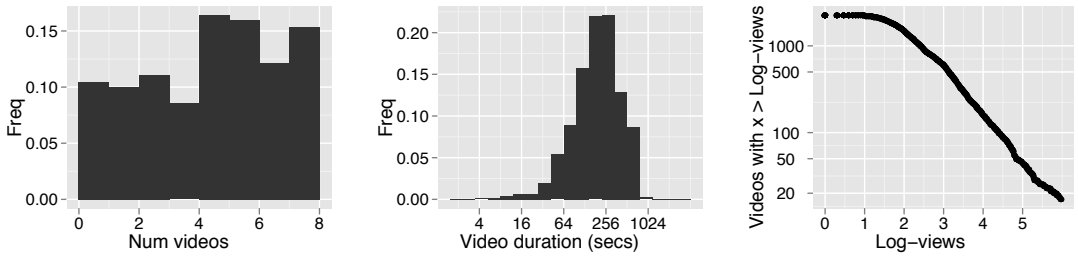

Figure 3.3: Some basic figures of the YouTube dataset: (left) the distribution of vlogs collected for the 469 users; (center) the histogram of the videos duration; (right) the cumulative distribution of views per video received in YouTube.

of vlogs in YouTube. Regarding gender, our sample shows a balanced distribution, with 47% of males and 53% of females.

3.3 Automatic processing of vlogs

Vlogs result in extremely diverse content and audiovisual quality, compared to purely conversational data recorded in controlled scenarios. 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 some 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.

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 3.4. 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 3.3.1 and 3.3.2, respectively.

3.3.1 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

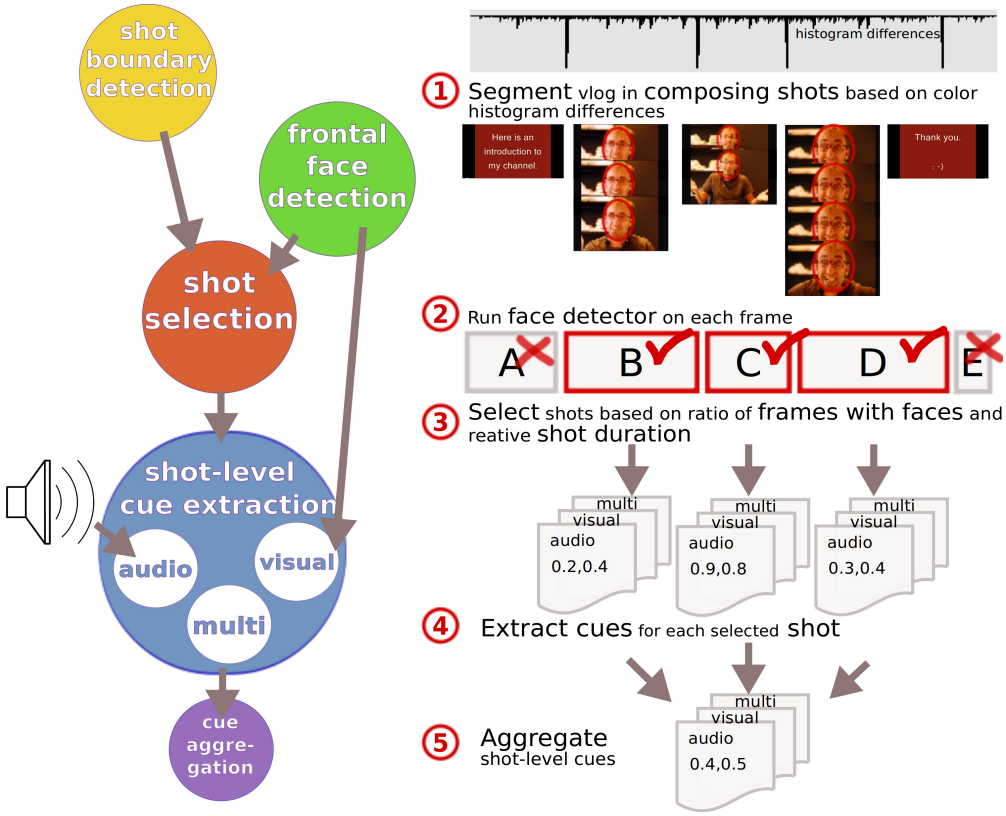

Figure 3.4: Automatic processing of vlogs: preprocessing (steps 1, 2, and 3) and nonverbal cue extraction (steps 4 and 5).

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, 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 face detections rate and the relative shot duration. 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. For the purpose of tuning shot boundary and conversational selection we annotated the shot discontinuities on a sample

of 100 vlogs (up to 168 hard shots), and labeled the shot conversational state ($s = 1$: conversational, $s = 0$: non-conversational). For shot boundary detection, we experimented with different thresholding methodologies proposed in the literature, including global, relative, and adaptive thresholds applied to histogram differences [Hanjalic, 2002], obtaining the best performance using a global threshold of $\gamma_g = 0.5$ ($EER = 15\%$ on the development set). Given the ratio of frames with faces in a shot $r_f = \frac{\text{\#of frames with faces}}{\text{\#video frames}}$ and the relative shot duration $r_d = \frac{\text{\#of shot frames}}{\text{\#of video frames}}$, conversational shots were classified by thresholding an estimate of the shot conversational state $\hat{s}$ computed using linear regression $\hat{s} = \alpha r_f + \beta r_d$ ($\alpha = 0.76$, $\beta = 0.24$, $R^2 = 0.6, p < 10^{-6}$). The best performance on the development set was obtained using a threshold $\gamma_c = 0.29$ ($EER = 7.5\%$).

3.3.2 Nonverbal Behavioral Cues Extraction

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 research [Pentland, 2008, Gatica-Perez, 2009]. Vocalic and motion cues have shown to be correlated with levels of interest, extraversion, and openness to experience, and are good predictors of dominance [Jayagopi et al., 2009], status [Jayagopi et al., 2008], and of 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 therefore we hypothesize that, to some extent, these cues may be suitable to characterize their behavior.

As shown in Figure 3.4, we first extract nonverbal cues for each conversational shot (step 4), and then aggregate cues from all shots to compute video features (step 5). In the next three subsections, we present the list of audio, visual, and multimodal cues we investigated. In addition, in the fourth subsection, we define a few features that we considered as potentially interesting to explore to characterize the level of editing in vlogs.

Audio cue extraction

We extracted speaking activity cues using the toolbox developed by the Human Dynamics group at MIT Media Lab [Pentland, 2008]. These cues are extracted on the basis of a two-level hidden Markov model (HMM) that is used to segment the audio in voiced/unvoiced and speech/non-speech regions (Figure 3.5, top). From there, several cues measure how talkative and fluent people are.

- The speaking time (*Speaking Time*) is a measure of how much the vlogger talks. Though our vlogs display a monologue setting, we hypothesize that some vloggers may be more talkative than others. This feature is computed by the ratio between the total duration

of speech and the total video duration.

- The length of the speaking segments (*Avg Length of Speak Segs*) is a measure of fluency, typically related to the duration and number of silent pauses (long segments are associated with short and few pauses) [Scherer, 1979]. It is measured by the ratio between the overall duration of speech divided by the number of speech segments.

- The number of speaking turns (*# Speech turns*) is another measure of fluency, directly related to the number of silent pauses (silent pauses interrupt and initiate speaking turns) [Pentland, 2008]. It is obtained by the ratio between the number of speech segments and the duration of the video.

In addition, the toolbox provides three prosodic cues:

- The voicing rate (*Voice rate*) relates to the number of phonemes produce while speaking, and represents the pace of a conversation [Scherer, 1979]. It is computed as the number of voicing segments divided by the total duration of speech segments.

- The speaking energy (*Energy*) is a measure of loudness, typically related to excitement. In its mean-scaled standard deviation form, it is also used to measure vocal control (how well the vlogger controls loudness), a feature typically related to emotionality [Pentland, 2008].

- The pitch (*F0*) is the main frequency of the audio signal. In its mean-scaled standard deviation form, it is another measure of vocal control and emotionality [Pentland, 2008].

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.

We use a smoothed version of the face detection output to construct a binary segmentation of intervals looking/non-looking at the camera and compute several cues related to looking patterns (see Figure 3.5). In addition, we use the face detection bounding box to measure proximity to the camera and framing. These features are described below:

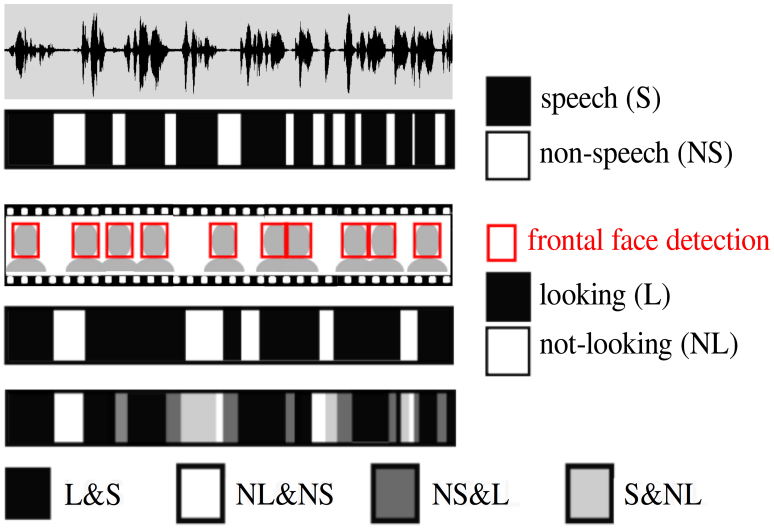

Figure 3.5: Nonverbal cues are extracted based on speech/non-speech, looking/non-looking segmentations, and multimodal segmentations. Looking/non-looking segmentations are based on frontal face detection. Multimodal segmentations are generated by combining speech/non-speech and looking/non-looking.

- The looking time (*Looking Time*) is a measure of how much the vlogger looks to the camera. We hypothesize that despite the clear communication intent of vlogs, vloggers may differ on the overall time spent looking at the camera. It is measured by the ratio between the overall looking time and the duration of the video.

- The length of the looking segments (*Av Length Look Seg*) is a measure of the persistence of a vlogger's gaze. It is computed by the ratio between the overall looking duration and the number of looking segments.

- The number of looking turns (*# Look turns*) measures how frequently the vlogger looks at the camera and it is directly related to patterns of gaze avoidance. It is obtained by the ratio between the number of looking segments and the duration video.

- The proximity to the camera (*Proximity to camera*) characterizes the choice of the vlogger to address the camera from a close distance. It is computed as the inverse of the average face bounding box area normalized by the video frame area.

- The vertical framing (*Vertical Framing*) measures to what extent faces are positioned on the upper part of the video frame and it is associated with vloggers showing the upper body. It is measured as the average vertical distance between the center of the bounding box and the center of the video frame normalized by the video frame height. A similar measure is used to compute horizontal framing (Hor Framing).

- *Head motion* is an indicator of excitement and kinetic expressiveness, together with

gestures. In this chapter, we computed an estimate of head motion obtained from difference between vertical (and horizontal) framing between consecutive frames.

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, 1982b] 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 moderate ratio [Dovidio and Ellyson, 1982b]. 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 previous subsections, we obtain a multimodal segmentation of speaking and looking patterns (see Figure 3.5). 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*.

Video editing cues

We explore the *number of shots per second* and the *video duration* (in seconds) as measures that may reflect vloggers' video editing practices. For example, largely 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.

3.4 Experiments and Results

This Section is organized as follows. In Section 3.4.1 we analyse vlogs to investigate the video edition elements involved in this genre. In Section 3.4.2, we provide an overview vlogger behavior on the basis on their aggregated nonverbal cues. Finally, in Section 3.4.3, we investigate the links between vlogger behavior and the social attention received by their videos.

3.4.1 Vloggers' Video Creation Practices

We performed a manual content analysis of vlogs with the purpose of understanding some of the video composition and edition practices involve on vlogging. Because this analysis is not feasible for the whole dataset, we limited it to a random sample of 100 vlogs and leverage the shot segmentation and conversation selection outputs to bring up several insights on some of these patterns at a larger-scale.

Component	Description	%
Snippets	Opening, ending or intermediate non-conversational video snippets	45
Opening	Preface video snippet used as a transition to the main conversational part of the vlog	16
Ending	Credits-like closing video snippet	20
Intermediate	Snippet that interrupts the conversational scene	35
Soundtrack	Music used in openings, endings, and intermediate snippets	25
Background music	Music used as background during the conversational scene	12
Object	Vloggers bring an object to the camera around which the conversation takes place	26

Table 3.1: Elements manually coded in a sample of 100 vlogs. The first six elements correspond to video editing elements. The metric % corresponds to the percentage of videos that contained at least one occurrence of the respective coded element.

Manual analysis

Table 3.1 summarizes the elements coded in our analysis and the percentage of videos featuring each one of the video editing elements. For each vlog, we coded elements as being present or absent in the vlog (independently of how many times they occurred). We adopted this coding mechanism as it was previously proposed in [Landry and Guzdial, 2008] to analyze the video composition elements in a more general sample of online video, and thus, allows for direct comparison.

The most popular composition element is the use of video snippets during video editing. 45% of the vlogs contained some kind of non-conversational snippet, which in most cases is found to be interrupting the main conversational scene, as shown by the 32% of the videos that contained at least one intermediate non-conversational snippet. During the coding process, we observed that these type of snippets typically correspond to video footage or edited slideshows of portraits, landscapes, and text, showing 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 use 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

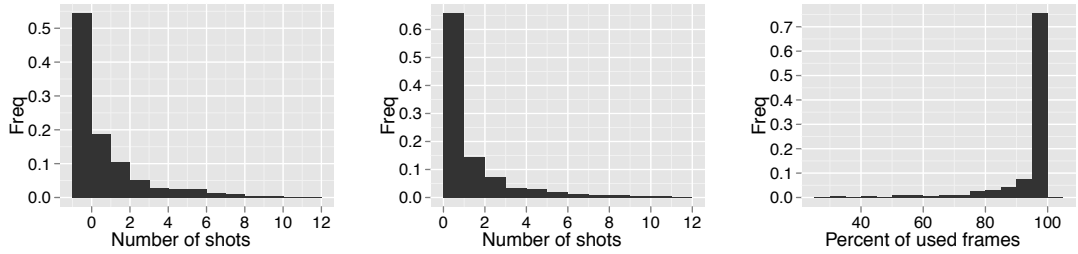

Figure 3.6: Automatic preprocessing output in terms of rejected and selected data. From left to right, distribution of non-conversational (rejected) shots, conversational (selected) shots, and percentage of frames selected per vlog. The large amount of data in terms of both shots and duration per vlog certifies the main conversational intend of vlogs.

poorly with respect to more general samples of video. In particular, the percentages of videos in our sample containing openings, closings, and soundtracks are about half of those found in a random sample of the 100 top YouTube popular videos analyzed by Landry and Guzdial [2008]. This suggests that vloggers focus mainly on the conversational aspect of the video, and exploit video editing to enrich and support their discourse.

Automatic analysis

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 in order to illustrate some characteristics of the whole dataset.

Figure 3.6 (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 3.6 (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 3.6 (right) shows the percentage of frames per vlog left after removing non-conversational shots, which indicates that the non-conversational content in vlogs 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

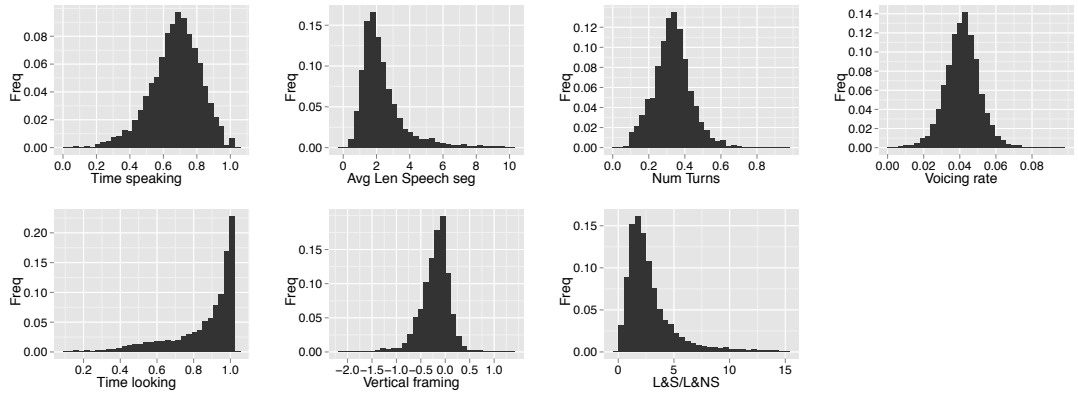

Figure 3.7: Selected nonverbal cue distributions for conversational shots in YouTube vlogs: four audio cues, three visual cues, and one multimodal.

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.

3.4.2 Vloggers' Nonverbal Behavior

We examined the automatic extraction of nonverbal cues by plotting their distributions and computing some basic statistics, both at the shot level and the video level. In this section, 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 of the same and different modality.

Figure 3.7 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 result from the fact that most of the vlogs are composed of few conversational shots, as shown in the previous section. 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 for more than half of the time, which suggests that vloggers who were perceived as mainly talking during the annotation done for the data collection (see Section 5.3.1) 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 that is similar to other conversational scenarios [Scherer,

		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										–

Table 3.2: Pearson's intra-feature correlations, ($^{*}\,p < .01, ^{**}\,p < .001, ^{***}\,p < .0001$).

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 the 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. 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, 1982b]. However, it may also be influenced by the fact that vloggers have nobody specific to look at, which in the case of long pauses could result in vloggers not starring the camera and thus having low L&NS.

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 3.2. 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, they are within the levels often reported in social psychology [Dovidio and Ellyson, 1982b, Ambady and Rosenthal, 1992].

3.4.3 Vloggers' Nonverbal Behavior and 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, 1982b, Borkenau and Liebler, 1992, Ambady and Rosenthal, 1992]. In social computing, works have shown that certain automatically extracted nonverbal cues in face-to-face interactions are related with perceived social attributes (e.g., dominance, role, attraction) [Pentland, 2008, Gatica-Perez, 2009]. 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 [Ambady and Rosenthal, 1992, Borkenau and Liebler, 1992, Ashton et al., 2002, Scherer, 1979].

A recent work investigating the phenomenon of attention in social media argued that social attention can be understood as a public reward to users who contribute with quality content [Huberman et al., 2009]. This work showed 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]. In this section, we hypothesize that, in the case of vlogging, part of the attention received by vloggers can be explained by their nonverbal behavior, as it happens in face-to-face interactions.

We investigate the correlation between automatic nonverbal behavioral cues and a measure of attention computed across populations of vloggers featuring similar behavioral cues. This aggregation method was borrowed from [Huberman et al., 2009], were it was used to aggregate the views received by users in a given time window. In our case, we use it to compensate views counts from network processes such as preferential attachment [Cha et al., 2007], which inflate these measures, but may not result from a clear interest of audiences in the content of the videos itself. Our correlation analysis is described as follows.

For each nonverbal cue:

1. Divide the numerical range of the cue into roughly L equally-sized sets of vloggers.

2. Let N_i be the number of videos in the i-th set $i = 1 \dots L$, and let $c_{n,i}$ and $v_{n,i}$ be the cue value and the view count corresponding to the n-th video in the set i, respectively, $n = 1 \dots N$. Then, for each $i = 1 \dots L$ compute the average nonverbal cue value :

$$\hat{c}_i = \sum_{n=1}^{N_i} c_{n,i} / N \tag{3.1}$$

   and the average social attention:

$$\hat{v}_i = \sum_{n=1}^{N_i} \log(v_{n,i}) / N \tag{3.2}$$

3. Compute the correlation between $\hat{c} = c_i$ and $\hat{v} = v_i$.

In practice, equations 3.1 and 3.2, can be replaced with other aggregate measures like the median. Table 3.3 shows the correlations between selected nonverbal cues (from audio, visual, and multimodal cues) and social attention, measured using $L = 50$ levels, whereas Figure 3.8 shows a selected number of these effects. For each video, nonverbal cues were aggregated from shots (step 5 in Figure 3.4) 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. 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 a number 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].

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				
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				
L&S	$.75^{***}$	$.79^{***}$	$.76^{***}$	$.73^{***}$
L&NS	$-.78^{***}$	$-.73^{***}$	$-.74^{***}$	$-.80^{***}$
L&S/L&NS	$.69^{***}$	$.68^{***}$	$.73^{***}$	$.68^{***}$

Table 3.3: 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).

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 videoconferences found significant differences between the empathy generated by people during showing the head compared to people showing the upper-body [Nguyen and Canny, 2009], due to the impossibility of the former setting to convey body language cues. The effect may be similar in vlogging, where the closer to the camera, the larger the face and smaller the proportion of upper-body shown. In effect, our vertical framing feature also shown larger levels of attention for heads positioned in the top side of the frame. Finally, the two measures of motion revealed a significant positive correlation with attention (up to $r = .70$, $r = .52$, $p < 10^{-3}$), concurring with results that associated successful

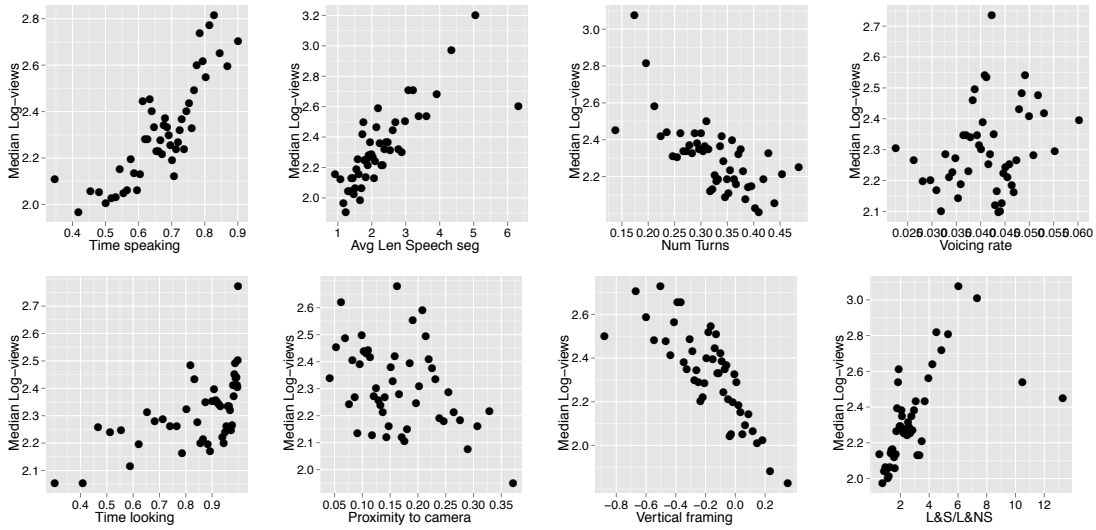

Figure 3.8: The aggregated nonverbal cues versus the social attention measure.

communication with visually active people [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), which suggest that vloggers displaying a dominant behavior are also associated to higher levels of attention.

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.

Accounting for the Temporal Dimension of Videos' Views

In the previous analysis, we used the accumulated view count of YouTube vlogs to measure social attention, without considering the date of upload of the video. While it is clear that older videos are publicly exposed longer time, we do not know the effect that this has neither on the view count, nor on our analysis. Thus, we investigate the use of a modified view count that compensates for the age of videos.

This analysis is similar to the on reported in [Huberman et al., 2009]. We divided the time span of our vlog uploads in $L = 50$ intervals corresponding to different dates of upload. Let N_i be the number of videos in i-th interval, and let $v_{n,i}$ and $\tau_{n,i}$ be the number of views and the age (number of days between the video's upload date and the data collection), of the n-th video in the i-th interval respectively, $n = 1 \ldots N_i$. Then, for each interval $i = 1 \ldots L$, we computed the average number of views $\hat{v}_i = \sum_{n=1}^{N_i} log(v_{n,i})/N$ and the average age $\hat{\tau}_i = \sum_{n=1}^{N_i} \tau_{n,i}/N$. Figure 3.9 shows the distribution of average views $\hat{v} = v_i$ with respect to the average times

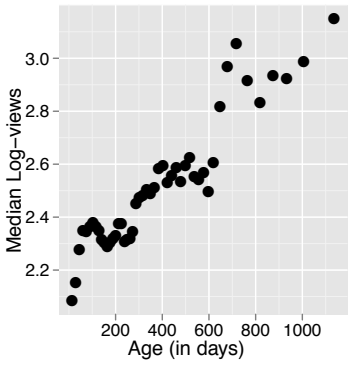

Figure 3.9: Social attention vs. age of videos (number of days between videos' upload date and data collection date). The average level of attention increases as the age of the video increases with very high correlation ($r = .94$, $p < 10^{-3}$).

$\hat{\tau} = \tau_i$ which evidences that older videos receive in average more views than recent uploads ($r = .94$, $p < 10^{-3}$).

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: $\hat{v}_i = a\tau + b$, where 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 computed an updated version of view counts. Consider a video in the interval i-th, let v be the original view count of the video, and let $\hat{v}_i$ be the average view count predicted for the interval i-th. Then, the updated view count v' is computed as $\log v' = \log v - \hat{v}_i$.

The experiments in social attention were replicated using the updated views. We found that accounting for the temporal dimension of the views did not result in significant differences between the correlation effects measured and the ones previously reported. Not surprisingly, this concurs with the negligible effect of videos' age in previous aggregate analysis of attention [Huberman et al., 2009].

3.5 Conclusions

In this chapter, we presented a framework to characterize nonverbal behavior in vlogs, based on the use of automatic audio and visual techniques that are robust to the variability of content found in this type of videos. In particular, we proposed a principled method to identify and discard the non-conversational parts of vlogs, by integrating shot-boundary and frontal face detectors, and then extracted nonverbal cues from the conversational parts.

In our experiments, we fist investigated the practices of vloggers in terms of video creation and edition. Our manual analysis of YouTube vlogs showed that **the use of these elements in vlogging is less frequent than that found in general samples of online video**. In addition, the analysis of both the automatic processing output of conversational vlogs and the distribution

of nonverbal behavioral cues confirmed the idea that content in this video genre is driven by a communicative intent.

A compelling result of our work is the evidence that **some audio, visual, and multimodal behavioral cues** extracted from vlogs 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 correlated with the average level of attention of their vlogs**. To the best of our knowledge, this is the first time that such a connection is found in online video.

We shall emphasize that we do not claim any direct causality between nonverbal cues and social attention. Rather, these results may provide initial evidence that, in addition to the topics addressed in vlogs, nonverbal behavior plays a role in the communication process of vlogging . These nonverbal cues are likely related to personal and social constructs such as personality traits (like extraversion) or persuasion, or to the impressions that audiences make of these characteristics, which is the problem investigated in the next chapters. We should also note that despite the high values of the correlations reported in this chapter, the effects explain the variance of the average level of attention received by videos displaying similar behaviors, and not the individual number of views of videos as it would result from a traditional correlation analysis.

As a final insight, we acknowledge the limitations of the automatic processing techniques proposed in this chapter. For example, both the shot boundary and the conversational shot detectors were optimized on a sample of 100 vlogs, which may seem small compared to the whole dataset. However, the set of 186 annotated shot boundaries compares with datasets used in works addressing shot boundary detection [Hanjalic, 2002]. In addition, analyzing the accuracy of the shot boundary detection was prohibitively expensive for the 150h of video in our dataset: in our experiments, the shot boundary detection method found 3,278 shot boundaries.

We also found evidence that some of the feature extraction methods have low cue validity, i.e., they do not entirely capture the aspect they intend to. For example, in the case of the looking/non-looking segmentation, low cue validity may be due to the simplification of using frontal face detections as indicating vloggers looking at the camera. Validating cues requires measuring nonverbal behavior manually, and is typically done in social psychology works with small data samples [Knapp and Hall, 2005], as it requires a substantial amount of time and expertise. On the other hand, research in social computing has shown that the use of simple, (although possibly inaccurate) 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. In contrast, these automatic nonverbal cues have high reliability [Curhan and Pentland, 2007], i.e. multiple runs of the same feature extraction algorithm in one video provide the same feature values, which together with cue validity is a desired requirement for behavioral measures in social psychology research.

4 Personality Impressions in Conversational Vlogging

4.1 Introduction

In this chapter, we address the problem of interpersonal impressions in vlogging from the perspective of personality research. We see the study of personality impressions, and in particular the use of the Big-Five factors model, as a natural starting point for the study of impressions in vlogging: first, because the model is useful to describe personal and social traits at a broad level, and second, because the abundance of social psychology literature in this topic can be used to contextualize research questions and findings. In addition, despite the absence of a clearly defined framework for the study of interpersonal impressions in social media, the study of the Big-Five traits has already been addressed in works studying impressions from personal websites [Vazire and Gosling, 2004], social network profiles [Gosling et al., 2007, Counts and Stecher, 2009], and text blogs [Li and Chignell, 2010].

The main goal of this chapter is to investigate how personality impressions are built based on the nonverbal and verbal behavior displayed by vloggers. As introduced in the previous chapter, the study of vlogger personality impressions can help to improve the understanding of why certain aspects of vlogger behavior influence social media attention. This problem is addressed in the first part of this chapter, where we use correlation analysis to investigate the connections between vlogger personality impressions and several social attention measures computed from YouTube metadata. The second part is dedicated to investigate the utilization of nonverbal and verbal behavioral cues for building personality impressions and to address the automatic prediction of personality impressions using machine learning algorithms.

Nonverbal behavior is effectively used to express aspects of identity such as age, occupation, culture, and personality, and therefore it is also used by people to form interpersonal impressions about other people during interactions [Ambady and Rosenthal, 1992]. Moreover, nonverbal cues are useful to characterize social constructs related to human internal states, traits, and relationships as shown in social psychology [Knapp and Hall, 2005] and social computing [Pentland, 2008, Gatica-Perez, 2009]. This is relevant in the context of social media, not only because nonverbal behavior has been unexplored in the context of personality

impressions, but also because the nonverbal channel conveys information that is often unconscious [Knapp and Hall, 2005] and more difficult to control than that of social network profiles and text blogs studied in previous personality works [Gosling et al., 2007]. In this chapter, we explore a large set of cues from audio and video, some of which had been previously used in the literature.

The human face has also been documented in social psychology literature as an important source of information in interpersonal impressions [Knapp and Hall, 2005, Knutson, 1996, Hall et al., 2011]. Kenny et al. [1992], for example, found evidence that facial cues such as smiling may contain information to form impressions of the Agreeableness trait that is not otherwise conveyed in other nonverbal cues. In addition, there is evidence that, among facial features, facial expressions of emotion provide information other than emotional states, influencing interpersonal impressions such as personality judgments, and that specific affective cues are in fact correlated with the possession of various personality traits [Knutson, 1996, Hall et al., 2011]. We argue that this may be specially true in conversational vlogging, where vloggers typically display head and shoulders on camera, and faces occupy a large portion of the screen. In this chapter, in addition to the audiovisual nonverbal cues mentioned above, we investigate cues derived from facial expressions as a potential source of information that the former features cannot convey. The analysis of facial expressions in webcam video has been recently researched in [McDuff et al., 2011] from the perspective of passive, mainly silent, viewing of advertising content, but to the best of our knowledge, our work is the first one to extract and study facial expressions in vlogs, where people are mainly talking.

Finally, we investigate the feasibility of using verbal content for the prediction of personality impressions from vloggers using both manual speech transcriptions and automatic speech recognition (ASR). Whereas the verbal channel is a clear alternative to the nonverbal channel and has already been investigated in social media through for the analysis of blogger personality [Gill et al., 2009a, Yarkoni, 2010, Li and Chignell, 2010], to the best of our knowledge, this is the first time verbal content is studied in the context of conversational social video. In addition, ASR technologies are the only means to truly scale verbal content analysis to the amount of online video available today. Recent work on the analysis of face-to-face conversations [Sanchez-Cortes et al., 2012] has already shown the potential of using ASR technologies to predict personal constructs.

We summarize the contributions of this chapter as follows:

- We conduct a correlation analysis to investigate the links between personality impressions and several measures of attention estimates from YouTube metadata. We show that impressions of the Big-Five associated to socially desired traits are useful to explain increasing levels of attention.

- We tackle the problem of individual cue utilization with standard pair-wise correlation analysis between automatic cues and personality impressions. We show that several

behavioral cues are associated to judgments of personality impressions that emerge from watching vlogs, and that some of these relations have been documented in social psychology literature.

- We address the task of automatically predicting vlogger personality impressions in social video, thus contributing to existing works on predicting personality and personality impressions in social media and social psychology research. We explore a broad set of nonverbal cues from voice and body, facial expressions and verbal cues as sources of personality information, and show that different modalities are useful to predict different traits with performance that compares to other personality prediction tasks reported in related literature.

- We present a first attempt to use a fully automatic framework for verbal content analysis in vlogs, where transcripts are obtained using automatic speech recognition. We evaluate the use of this output for the task of personality prediction, and compare it to manual transcriptions.

Some of the work presented in this chapter has been previously published. The automatic analysis of audiovisual nonverbal cues for personality prediction appeared in [Biel et al., 2011] and [Biel and Gatica-Perez, 2012b]. The use of facial expressions of emotion appeared in [Biel et al., 2012]. The parts on automatic verbal content analysis and fusion of features are unpublished.

The rest of this chapter is structured as follows. In Section 4.2, we introduce the Big-Five model of personality. In Section 4.3, we describe our dataset, including the personality annotations and the manual and automatic transcriptions. We introduce the feature extraction processes in Section 4.4. We describe our experiments on the task of personality modeling in Section 4.5, and present all the results in Section 4.6. Finally, we conclude in Section 4.7.

4.2 The Big-Five Personality Model

The Big-Five framework of personality is a hierarchical model that organizes personality traits in terms of five basic bipolar dimensions: Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to Experience (O). These five dimensions are easily interpreted by referring to their associated personality attributes, as presented in Table 4.1 [McCrae and John., 1992].

Though the Big-Five model has not been universally accepted, it has considerable support and has become the most widely used and researched model of personality [Gosling et al., 2003]. To measure the extent to which each one of these traits describes human personality, several rating instruments have been developed and used to ask people to rate themselves

Personaliy Trait	Adjectives
Extraversion (E)	Active, Assertive, Energetic, Enthusiastic, Outgoing, Talkative
Agreeableness (A)	Appreciative, Forgiving, Generous, Kind, Sympathetic
Conscientiousness (C)	Efficient, Organized, Planful, Reliable, Responsible, Thorough
Neuroticism* (N)	Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying
Openness to Experience (O)	Artistic, Curious, Imaginative, Insightful, Original, Wide Interests

Table 4.1: Big Five personality traits and associated adjectives [McCrae and John., 1992]. *Neuroticism may alternatively be presented as Emotional Stability by inverting the scale.

(a.k.a. self-reported personality) or to rate others (a.k.a. personality impressions) based on a set of questions or items. Self-reported personality and personality impressions provide two different views of personality perception and have been used to answer questions such as: a) do observers agree on their personality impressions? b) are personality impressions accurate compared to self-reported personality? c) How much information is needed to achieve agreement (or accuracy) in personality impressions?

Regarding a) and b), research has shown that external observers agree on their personality impressions of targets, and that such impressions are fairly accurate with the targets' self-reported personality, even if impressions are formed in the presence of minimal information, in several contexts: physical presence [Ambady et al., 1995], video-tapes [Borkenau and Liebler, 1992, Kenny et al., 1992], photos [Borkenau and Liebler, 1992], websites [Vazire and Gosling, 2004], bedrooms [Gosling et al., 2002], etc. Regarding c), research has investigated the extent to which certain appearance or behavioral cues convey information associated to personality impressions (cue utilization) and to the actual self-reported personality (cue validity), and has shown that valid and useful cues vary depending on the the context in which impressions are made [Gosling et al., 2002, Vazire and Gosling, 2004]. In this thesis, we investigate the problem of cue utilization in vlogging.

Substantial work developing personality inventories has focused on ensuring that instruments are valid (i.e., a test consistently measure what it intends to measure) and reliable (i.e., a test produce similar results if repeated under consistent conditions). Many personality questionnaires are lengthy and take time to respond. For example, the BFI (Big Five Inventory, 44 items), the NEO-FFI (Neuroticism-Extroversion-Openness Five Factory Inventory, 60 items), and the TDA (Trait Descriptive Adjectives, 100 items) take approximately 5, 15, and 15 min to complete respectively [John and Srivastava, 1999]. Recent research has also developed briefer measures such as the TIPI (Ten-Item Personality Inventory), that despite being less reliable than the BFI, NEO-FFI, or TDA, are convenient when time is limited [Gosling et al., 2003]. In our work, the use of a short test is critical for the feasibility of large-scale analysis and the

	Trait	Mean	SD	Skew	Min	Max	1	2	3	4	5	$ICC(1,K)$
1	Extr	4.61	1.00	−.32	1.90	6.60		.03	.00	.08	$.56^{***}$	$.76^{***}$
2	Agr	4.68	.87	−.72	2.00	6.50			$.39^{***}$	$.69^{***}$	$.28^{***}$	$.64^{***}$
3	Cons	4.48	.78	−.32	1.90	6.20				$.55^{***}$	$.26^{***}$	$.45^{***}$
4	Emot	4.76	.79	−.57	2.20	6.50					$.31^{***}$	$.42^{***}$
5	Open	4.66	.71	−.09	2.40	6.30						$.47^{***}$

Table 4.2: Descriptive statistics, pair-wise correlations, and Intraclass Correlation Coefficients for Personality Impressions (ICC(1,k)), $^{***}p < .0001$.

simplicity expected for tasks suitable to be crowdsourced.

4.3 The YouTube Vloggers Personality Dataset

The dataset used in this chapter consists of 442 one-minute vlogs slices extracted from our YouTube vlog dataset; a set of personality impressions from vloggers, and the manual and automatic transcription vloggers' speech.

The video dataset comprises most of the individuals in our original collection in Chapter 3 as each vlog corresponds to a different vlogger, including 208 males (47%) and 234 females (53%). The one-minute slices correspond to the first conversational minute obtained using the preprocessing method explained in Chapter 3, and was used to limit the task of personality impression annotation to thin-slices of vlogs. The data is described in detail below.

Personality Impressions

The personality impressions determine to what extent each of the 442 vloggers can be described on the basis of the Big-Five traits, and were collected using Amazon's Mechanical Turk and the TIPI questionnaire. The personality scores of each vlogger result from averaging the scores given by five different MTurk workers after watching the one-minute vlog. Table 5.3 shows some descriptive statistics, pair-wise correlations, and Intraclass Correlation Coefficients (ICC(1,k)) of these impressions. The ICC(1,k) reliability of the aggregates indicates that MTurk worker judgments are comparable to those obtained in other settings: Extraversion (Extr, .76), Agreeableness (Agr, .64) Conscientiousness (Cons, .45), Emotional Stability (Emot, .42), Openness to Experience (Open, .47). While details on the crowdsourcing process and reliability of these annotations are discussed in detail in Chapter 5, here we summarize three main aspects that are useful to follow the discussion throughout this chapter: a) different personality traits achieve substantially different agreement; b) as in most literature, Extraversion is the trait achieving the highest agreement; and c) compared to other settings, the high reliability of Agreeableness seems particular of the vlogging.

Word Recognition Performance	
Correct (C)	45.6%
Substitution (S)	28.5%
Deletions (D)	25.9%
Insertions (I)	8.0%
Total Error (WER)	62.4%

Table 4.3: Word Recognition Performance for automatic transcriptions in 397 vlogs compared to automatically aligned manual transcriptions. $WER = \frac{S+D+I}{S+D+C}$.

Manual Transcriptions

We use a professional company to manually transcribe the audio from vlogs. However, whereas the personality impressions were annotated based on the one-minute vlogs, the data sent for annotation corresponded to the full vlog duration from 426 randomly selected vloggers, up to a total of 30h of audio. 5% of the vloggers were later discarded during the transcription process because their audio was unintelligible or they were not speaking English.

The resulting manual transcripts comprise 408 vloggers out of the 442 with personality impressions, and they are also balanced in gender: 197 males (48%) and 211 females (52%). The whole transcriptions contained a total of 10K unique words and 243K word tokens, and do not include timestamps.

Automatic Transcriptions

We used a state-of-the-art automatic speech recognizer provided by Dr. John Dines (Idiap Research Institute) [Hain et al., 2012] to generate transcriptions for the audio files and to align the manual transcriptions (used later for evaluation). The system combines two two-pass English systems that use acoustic models based on individual head-mounted microphones (IHM) and single-distant microphone (SDM), respectively. Both the IHM and SDM systems use identical decoding configurations, e.g., the first pass uses unadapted acoustic models, followed unsupervised adaptation in the second pass. The four hypotheses from both the first and second passes of IHM and MDM are aligned and combined to produce word-level confidence scores, and decoded with a weighted finite-state transducer using a lexicon of 50,000 words and a 4-gram language model trained on various corpora for a total amount of about one billion words.

The system was run on the 408 vlogs with manual transcriptions, but failed in up to 11 vlogs during alignment or decoding mostly because of the low audio quality and background noise. As summarized in Table 4.3, the ASR system achieved a word error rate (WER) of 62.4%, with respect to the aligned manual transcriptions for the 397 successfully decoded files. Unfortunately, we do not know what percentage of errors may have been caused by misalignments, as manual transcriptions do not include timestamps.

Though these results may seem low compared to the WER achieved in other domains, they concur well with another recent work that automatically transcribed YouTube videos [Hinton et al., 2012]. In those experiments, an ASR system trained on acoustic models from 1,400h of aligned YouTube audio, incremented by more than double the WER achieved in more controlled datasets, up to a WER of 52.3%. These results clearly illustrate the current difficulty of automatically transcribing online social video.

4.4 Automatic Behavioral Feature Extraction

In this chapter, we investigate the use of three different sources of personality information from vloggers. First, we automatically processed the one-minute vlogs to extract nonverbal cues from audio and video. The number of cues extracted was incremented with respect to Chapter 3 with the addition of more prosodic cues and cues related to audiovisual activity. Second, we propose an approach to extract cues from vlogger facial activity using the frame-by-frame estimates of a facial expression recognition system. Finally, we propose the extraction of word usage features from manual and automatic speech transcriptions of the vlogger speech.

4.4.1 Audiovisual Nonverbal Cues

We automatically processed the one-minute vlogs to extract nonverbal cues from audio and video.

Acoustic Activity Nonverbal Cues

We computed the nonverbal cues measuring acoustic activity using the MIT Media Lab audio toolbox. The description of these cues can be found in Section 3.3.2 and includes the speaking time (Speaking Time), the average length of speaking segments (Av Len Speak), and the number of speaking turns (# Speech Turns).

Prosodic Cues

Several prosodic cues were also obtained from audio. Voicing rate and some related cues were obtained from the MIT audio toolbox in a frame-by-frame basis (with windows of 32ms and steps of 16ms). In addition, energy and pitch features were obtained from the audio signal using PRAAT in windows of 40ms and time steps of 10ms. The mean-scaled standard deviation energy and pitch were already used in previous chapter extracted using the MIT toolbox. Here, we aggregated energy and pitch across frames and computed the mean, median, mean-scaled standard deviation, maximum, minimum, and entropy.

- The voicing rate (*Voice rate*) relates to the frequency of phonemes while speaking, and represents the pace of a conversation [Scherer, 1979]. In monologues, it can represent

a measure of fluency or excitement. It is computed by the ratio between the overall duration of voice segments and the duration of speech. In addition to the voicing rate used in previous chapter, we use the number of autocorrelation peaks (*# R0 peaks*), their location (*Loc R0 peaks*) and the spectral entropy (*Spec Entropy*) which are the raw features used to obtain the voicing/non-voicing segmentation [Basu, 2002].

- The speaking energy (*Energy*) is a measure of loudness, typically related to excitement. In this chapter we use the mean, median, mean-scaled standard deviation, maximum, minimum, and entropy of energy and first derivative of the Energy (*D Energy*).
- The pitch (*F0*) is the main frequency of the audio signal. In its mean-scaled standard deviation form, it provides another measure of vocal control and emotionality [Pentland, 2008]. In addition to the pitch, we obtained the pitch bandwidth (*F0 BW*), intensity, (*F0 Intensity*), and the confidence of the estimate (*F0 Conf*), which were respectively aggregated as mentioned above.

Looking Activity and Pose Cues

We computed looking activity cues on the basis of looking-non-looking segmentations from frontal face detections, including time looking at the camera (Looking Time), average length of looking segments (Av Len Look Seg). Amongst the pose cues proposed in the previous chapter, we choose proximity to the camera (Proximity to Camera), and vertical framing (Vertical Framing) because they were the two ones that showed larger relation to social attention (see previous Section 3.3.2 for a description of these cues and Section 3.4.3 for the social attention analysis).

Visual Activity Cues

In collaboration with Dr. Oya Aran (Idiap Research Institute), we automatically extracted a set of visual cues as descriptors of the overall visual activity of the vlogger throughout the video using a modified version of motion energy images Bobick and Davis [2001], that we call "Weighted Motion Energy Images" (wMEI). The wMEI is calculated as:

$$wMEI(x,y) = \frac{1}{N}\sum_{t=0}^{T} D(x,y,t), \tag{4.1}$$

where $D(x,y,t)$ is a binary image that shows the moving pixels (x,y) time t, N is the normalization factor (see below), and T is the total number of frames. Figure 4.1 shows some examples of wMEI images. Unlike motion energy images, wMEI is not a binary image. In wMEI, the brighter pixels correspond to regions where there is more motion.

A wMEI is normalized by dividing all the pixel values by the maximum pixel value. Thus, a normalized wMEI contains the accumulated motion through the video as a gray-scale image, where each pixel's intensity indicates the visual activity in the pixel (brighter pixels

Figure 4.1: Example of wMEI images computed for different vloggers. Bright pixels correspond to regions with more motion.

correspond to regions with higher motion). From the normalized wMEIs, we extract simple statistical features as descriptors of the vlogger body activity such as the entropy (wMEI e), mean (wMEI m), median (wMEI md), and center of mass in horizontal (wMEI H Cog) and vertical dimensions (wmEI V Cog). To compensate for different video sizes, all images are previously resized to 320x240 pixels.

Multimodal cues

We also extracted multimodal cues from speech/non-speech and looking/not-looking segmentations. These cues are described in detail in Section 3.3.2, and include: of looking-while-speaking (L&S), looking-while-not-speaking (L&NS), and the multimodal ratio (L&S/L&NS).

4.4.2 Facial Expression Cues

In collaboration with Lucia Teijeiro (University of Vigo, Spain), we investigated several methods to vlogger facial expressions of emotion using the output of a facial expression recognizer.

Facial expression analysis has been thoroughly researched during the last two decades [Fasel and Luettin, 2003, Donato et al., 1999] in the computer vision field. The Facial Action Coding System (FACS) developed by Donato et al. [1999] has become a standard framework for detecting facial actions and for classifying facial expressions of emotion. FACS defines the action units (AUs) that code the movement of facial muscles, and are considered as the fundamental units of facial expressions. The FACS system has been used to uniquely define seven universal facial expressions of emotion in terms of AUs: Anger, Contempt, Surprise, Fear, Joy, Sad and Disgust.

In our work, we processed the one-minute-slices using the Computer Expression Recognition Toolbox (CERT) [Littlewort et al., 2011], which is a real-time face processing software developed for facial expression understanding and constitutes a state-of-the-art in the field. CERT combines three face processing stages: face detection, feature detection, and face registration,

to obtain a cropped face patch used for expression analysis. Based on the AUs estimated using Gabor-based filters and SVM classifiers, CERT uses a multivariate logistic regressor to predict the seven expressions of emotion plus a neutral expression.

First, we preprocessed vlogs to detect the face, eyes, nose, and mouth of vloggers with a Viola-Jones face and facial feature detector [Agulla et al., 2009], and selected the videos in which most of the frames showed all facial features. After selection, the dataset was reduced to 281 vloggers (with the same gender distribution). This method was introduced with the purpose of minimizing errors in the face registration step previous to the facial expression extraction, which is critical given the diversity of vlogger poses and camera position settings in YouTube vlogs.

Then, we proposed a systematic method to model the frame-by-frame facial expression outputs from CERT into aggregate cues that capture different activation patterns from the seven facial expression and smile. First, we converted the CERT output to a binary segmentation that divides the expressions into active/inactive regions using two different approaches:

- Thresholding (THR): we consider the CERT output to be activated when frame values are larger than a threshold λ. For facial emotions we choose $\lambda = .005$, which represents a rather conservative choice, whereas for smile the threshold was set up to $\lambda = 0$ by definition of the CERT output [Littlewort et al., 2011].
- HMM: we implemented a two-state (active/inactive) Hidden Markov Model to detect the active state of the CERT output. Each output is modeled with one single Gaussian initialized with the threshold-based segmentation, and the transition probabilities are set to $\rho_{00} = \rho_{11} = .95$ and $\rho_{01} = \rho_{10}$.

In practice, the THR approach copes with high frequency changes, gives shorter and more frequent active states, whereas the HMM provides a smooth version of the output generated by CERT. Let r_i, $i = 1 \dots N_r$ be the activation state of the $i-$th region of the segmentation ($r_i = 1$ if active, $r_i = 0$ if inactive), N_r the total number of regions, N the total number of frames in the video, and f the video frame rate. Then, we define four different facial cues to measure the presence of expressions, their frequency, and their duration:

- Proportion of active time (PT): computed as $PT = \frac{1}{N}\sum_{i=1}^{N_r} \tau_i r_i$, where τ_i is the duration of region i (in frames).
- Number of active segments (NS): computed as $NS = \frac{1}{Nf}\sum_{i=1}^{N_r} r_i$ (i.e, the number of active segments per second).
- Average duration of the expressions (AD): computed as $AD = \frac{1}{N_r f}\sum_{i=1}^{N_r} \tau_i r_i$ (in seconds).
- Proportion of short segments time (PTS): computed as $PTS = \frac{1}{N_r}\sum_{i=1}^{N} \tau_i r_i$, $\forall \tau_i \leq .001 f$, i.e., the proportion of time in segments shorter than 100ms. This activity cue was introduced to explore the characterization of facial expressions as signals of short duration.

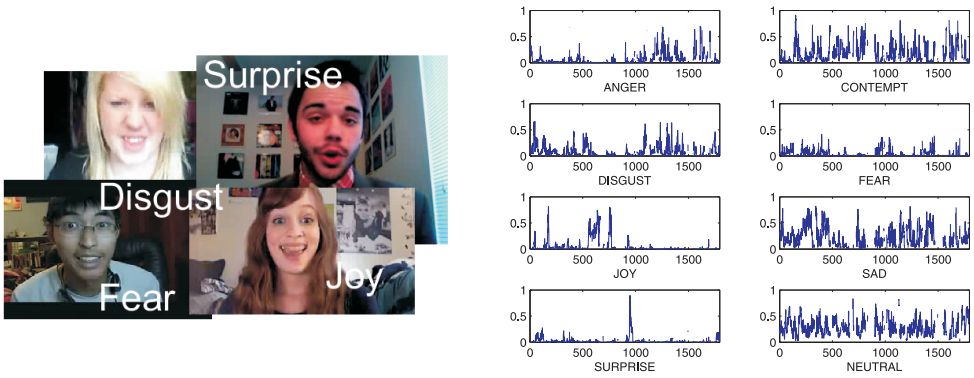

Figure 4.2: Left: example of facial expressions of emotions extracted from our YouTube dataset. Right: example of seven universal emotion signals plus neutral output by CERT.

The THR and HMM-based cues, were complemented with a third set of basic statistical features (STATS), computed directly on each facial expression signal. This set included the mean, median, variance, maximum, minimum, and entropy values.

4.4.3 Verbal Content Cues

Research in social psychology has shown the words that people use in their daily interactions reflect information about their personality [Pennebaker and King, 1999]. In this section, we explore two different methods used in the social media literature to compute verbal cues from manual and automatic transcriptions of vlogs. This was done in collaboration with Vagia Vtsiminaki (EPFL).

LIWC

The Linguistic Inquiry and Word Count software [Pennebaker and King, 1999] is built on a dictionary composed of 4,500 words and word stems that are classified into linguistic and paralinguistic categories, and other categories related to psychological constructs and personal concerns. For every document, the LIWC output is a breakdown of word category usage based on relative word occurrences, i.e., the number of occurrences of all words within that category is divided by the total number of words in a text (note that in LIWC, words can be assigned to more than one category at a time). In this work, we consider a total of 65 LIWC cues: we discarded the 12 punctuation categories, as there are not relevant in the spoken setting, and we included only one general descriptor that counts the words that are longer than six letters.

N-grams

Social media research showed that a n-gram approach to model verbal content with a proper selection procedure could outperform LIWC for the task of automatically predicting blogger

	Manual				Automatic	
	words	LIWC	uni	bi	words	LIWC
# terms	10K	65	1K	287	7,6K	65
# tokens	246K	221K	241K	110K	152K	142K

Table 4.4: Number of unique terms and tokens in manual and automatic data: raw vocabulary (words) and data processed using LIWC and n-grams (uni, bi).

personality [Iacobelli et al., 2011]. The n-gram model is a standard characterization of text used in many tasks related to text-based retrieval and classification of documents. In particular, this worked proposed the use of Weka's Correlation-based Feature Subset Selection (CFS) to select significant n-grams prior to the prediction experiments (a different subset of n-grams is used for each personality trait) and then to train a machine learning classifier.

In our work, features consist of $tf \cdot idf$ values of unigrams and bigrams including stop words, one of the representations proposed in Iacobelli et al. [2011]. However, we believe that the selection process proposed in Iacobelli et al. [2011] may be prone to overfitting, and therefore, we consider applying CFS in two different settings to contrast the results. In the first setting, we use CFS inside the evaluation set up (during training), whereas in the second, we used CFS outside the evaluation setup, as in [Iacobelli et al., 2011]. For the rest of the chapter, we refer to them as inCFS and outCFS respectively. Prior to generating n-grams, we preprocessed text by stemming words using Porter's stemming algorithm, removing punctuation, and omitting words that appeared in less than ten documents or less than ten times in the whole collection.

Table 4.4 summarizes the amount of data for manual (M) and automatic (A) transcriptions, including raw data (words), and the LIWC and n-gram outputs. 91.7 % of the words from manual transcripts were found in the LIWC dictionary, whereas for automatic transcripts this percentage decreased to 66% . The actual feature sets for uni and bi-grams are much smaller after using inCFS and outCFS (in most cases included no more than 100 features).

4.5 Personality Prediction Models

We treat personality inference as five independent regression problems intended to predict the personality scores for each of the Big-Five impressions. Compared to other prediction tasks proposed in the literature such as personality classification or ranking, the regression task is the one that provides the most fine-grained personality recognition assessment [Mairesse et al., 2007]. Our goal in this chapter is to assess what level of prediction performance can be achieved when using cues together with machine learning techniques. In addition, we also aim to evaluate to what extent we can make accurate predictions on the basis of moderately reliable crowdsourced annotations.

For each task, we evaluated the use of two different supervised machine learning predictors: Support Vector Machines (SVMs) with linear, polynomial, and RBF kernels, and Random

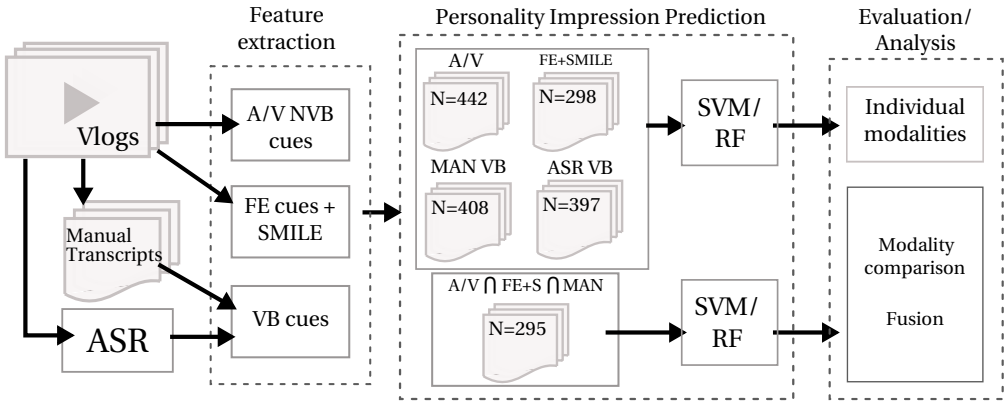

Figure 4.3: Our approach to study personality impression predictions. MAN = manual, A/V NVB = audiovisual nonverbal cues, FE cues = facial expression cues, VB cues = verbal cues ASR VB = verbal content from ASR, MAN VB = verbal content from manual transcripts, N = size of dataset after feature extraction.

Forests (RFs). We conducted experiments using a 10-fold cross-validation (CV): at each resample iteration, we train a model using 9 folds of data, and test it on the left-out fold. To optimize the model parameters, we used 5-fold cross validation on the 9 folds used for training. Note that we use CV with RF for practical reasons but that results were the same using the out-of-bag estimates of RF (which are performance estimates computed on bootstrap left-out data). Whereas the linear kernel consistently underperformed the RBF and the polynomial kernel, the performance of the RBF and the polynomial kernel was almost the same for all the tasks (only in few cases the RBF provided slightly better performance than the polynomial). Hence, to keep the presentation of the results clear, we decided to only report performance for the SVMs using RBF kernel, and for RFs.

Figure 4.3 shows a summary of our experimental protocol for the task of personality prediction (note that this does not include our experiments on social attention and correlation analysis). Following the previous section, different data and feature extraction procedures generated feature sets of different sizes. Our first experiments make use of the totality of data for each set, to investigate the performance of individual modalities. Then, we focus on the subsample of vloggers that intersect with all feature sets to compare the performance across modalities, and to explore the fusion of these three different sources. Because of the limitations of data, our fusion strategy consisted on concatenating feature vectors together to train one single model.

In all the experiments, we measured the performance of the automatic predictions using the coefficient of determination (R^2). The R^2 is measured based on the ratio between the model's absolute error and a baseline regressor that predicts the mean personality score (MPS). Formally, it is expressed as:

	Measure	*Mean*	*SD*	*Skew*	*Min*	*Max*	1	2	3	4	5
1	# Views	288.91	8.18	.98	1	2406284		.86	.88	.85	−.28
2	# Favorited	2.07	3.66	1.93	0	22007			.90	.86	−.18.
3	# Raters	14.71	4.71	1.27	1	27349				.92	−.20
4	# Comments	12.50	5.31	.99	0	23112					−.17
5	Average rating	4.85	.87	−.76	1	5					

Table 4.5: Descriptive statistics and pair-wise correlations for YouTube attention measures (all correlations are significant with $p < .0001$).

$$R^2 = 100 \times \left(1 - \frac{\sum (y_{obs} - y_{pred})^2}{\sum (y_{obs} - \bar{y}_{obs})^2}\right), \tag{4.2}$$

where y_{obs} and $\bar{y}_{obs}$ are the observed scores and their mean, respectively, and y_{pred} are the scores predicted by the model.

4.6 Results and Discussion

This section is divided in three parts. In Section 4.6.1, we investigate the links between personality impressions and social media attention as measured from the YouTube vlog metadata. In Section 4.6.2, we measure the utilization of automatic extracted cues from vloggers as lenses that mediate the personality impressions of observers by means of pair-wise correlations. Finally, in Section 4.6.3, we investigate the task of automatically predicting personality from vlogger behavior and content.

4.6.1 Personality Impressions and Social Attention

Several forms of social participation take place around vlogs. After watching a vlog, people may actively "like" or "unlike" it – manifesting some kind of approval or disapproval –, mark it as one of their favorites, or write a comment about it. All these actions, which are registered and aggregated by the YouTube platform in the form of metadata, can be interpreted as signaling the attention that vloggers receive from the YouTube audience. Because each one of this actions involves people to a different degree, one could argue that these different actions unveil different aspects of people's attention. For example, anybody can watch videos in YouTube, but only registered and logged-in people can "like", "favorite", or comment in a vlog. Moreover, writing a comment clearly takes more time and effort than just "liking" the video.

Table 5.6 summarizes some basic statistics of these measures as obtained from YouTube metadata for the 442 vlogs in our dataset. The values show that the number of views (# views),

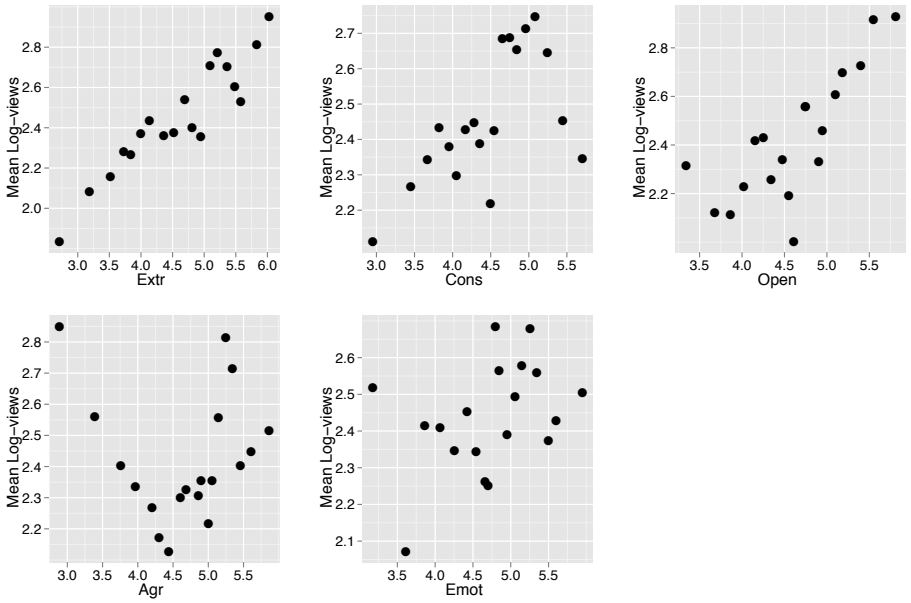

Figure 4.4: XYplots for personality impressions and social attention based on # views received. Relations for Extraversion, Conscientiousness, and Openness to Experience are mostly linear, whereas for Agreeableness is U-shaped.

times favorited (# favorites), raters (# raters) and comments (# comments) are highly skewed to very large values, due to the long-tail distribution of social network data, where a small percentage of vloggers are very popular and lots of them are not ordinary. To reduce this skewness, we transformed the measures with a log function. Note that the mean, standard deviation, minimum, and maximum values shown in the table were computed in the logarithmic scale and were transformed back for displaying purposes. Only the skewness measure was computed after transformation. Compared to the rest of the measures, the distribution of the average rating showed a large bias towards large values (mean = 4.85), which suggests that when people decide to rate a vlog, they tend to give high ratings. We transformed this measure with a power ten function to reduce the negative skewness of this measure.

We also computed the inter-class correlations between these measures, which are shown in the right part of Table 5.6. These values show, that despite the possible different nature of these participation measures, most of them are largely correlated. Note that only the correlations on the last column are relatively low, which indicates that the average rating received by the videos is weakly related to how much the vlog was viewed, favored, rated or commented.

We investigate the personality correlates with aggregates of these measures of attention computed across groups of vlogs featuring similar personality scores. This aggregation method was already introduced in Chapter 3 to analyze the link between automatic nonverbal cues and view counts; here, we apply it to the personality scores and other measures of attention.

For each personality trait and for each measure of attention:

Measures	*E*	*C*	*O*	*A*	*ES*
# Views	$.93^{***}$	$.61^{*}$	$.78^{***}$	$.70^{*}$	.36
# Times favorited	$.95^{***}$	$.58^{*}$	$.82^{***}$	$-.67^{*}$	$-.09$
# Raters	$.96^{***}$	$.51^{\dagger}$	$.82^{\dagger}$	$.65^{\dagger}$	$-.09$
# Comments	$.97^{***}$	$.61^{*}$	$.78^{***}$	$.60^{\dagger}$	.00
Average rating	.12	.26	$.67^{*}$	$.56^{\dagger}$	.24

Table 4.6: Pearson's correlation coefficients between vlog attention measures and personality impressions ($^{\dagger}p < .05, ^{*}p < .01, ^{**}p < .001, ^{***}p < .0001$). For Agreeableness, the correlation score results from a square relationship.

1. Divide the personality score range into roughly L equally-sized sets of vloggers.

2. Let N_i be the number of videos in the i-th set, $i = 1 \ldots L$, let $s_{n,i}$ and $a_{n,i}$ be the personality score and the attention count corresponding the n-th vlogger in the i-th set, respectively. Then, for each $i = 1 \ldots L$ compute the average personality score:

$$\hat{s_i} = \sum_{n=1}^{N_i} s_{n,i} / N \tag{4.3}$$

   and compute the average social attention:

$$\hat{s_i} = \sum_{n=1}^{N_i} a_{n,i} / N \tag{4.4}$$

3. Compute correlation between $\hat{a} = \{\hat{a_i}\}$ and $\hat{s} = \{\hat{s_i}\}$.

For our analysis, we used $L = 20$, as a compromise between the number of videos per set and the number of data points for the analysis.

Figure 4.4 shows the xyplots of the Big Five personality impressions and the attention measure using the number of views. The figure features a linear association between Extraversion, Openness to Experience, and Conscientiousness with attention, suggesting that users perceived as high scorers for these traits receive a higher level of attention from the audiences. Intuitively, it is reasonable to think that vloggers perceived as more extraverted or opened to experience may be more appealing or interesting to watch, because of the ways in which they create their videos and behave in them (including both the verbal and nonverbal channels). In addition, these type of personalities are more likely to be active and socially involved online as well as offline, and therefore might ultimately be recognized by the vlogger community. In contrast, Agreeableness shows a nonlinear association with attention, suggesting that the "pleasant" and "disagreeable" vloggers in the extremes of this personality dimension tend to receive more attention.

We measured the strength of all the possible associations between measures of attention

and personality by means of linear fits with the exception of those involving the Agreeableness impressions. For Agreeableness, we fit a second-order polynomial with all measures of attention except with the average rating, for which the association was also observed to be linear. Table 4.6 summarizes the correlation coefficient for all the fits (the R^2 values of the second order polynomial fits were converted to correlation coefficients for consistency using a square root, $r = \sqrt{R^2}$). The linear and nonlinear associations observed for the number of views in Figure 4.4 are also measured for the number of times favorited, comments, and raters, whereas the Agreeableness polynomial fit compares in strength to the rest of the fits. The case of average rating is interesting because it shows low correlation with Extraversion and Conscientiousness, suggesting that low and high extraverts are both likely to have high average ratings. On the contrary, the Agreeableness and Openness to Experience impressions are linearly associated to the average rating, suggesting that users scoring high in these traits are also obtaining higher average ratings. Thus, the "liking" social function associated to the average rating seems to capture well the meaning of agreeableness, giving lower ratings to disagreeable vloggers and higher ratings to pleasant, likable vloggers.

To sum up, the personality impressions align strongly with the audience response as measured from the aggregation of YouTube's metadata. From our view point, these results indicate that, in vlogging, the phenomenon of social attention can be explained to some extent by similar interpersonal perception processes to those occurring in face-to-face interaction.

4.6.2 Behavioral Cues and Personality Impressions

We investigated the individual correlations between automatic nonverbal behavioral cues and personality impressions by means of pair-wise correlations. This is a common approach of research in personality, mostly in social psychology [Scherer, 1979, Gosling et al., 2002] used to explore what aspects of targets observers may have used to make their judgements (i.e., cue utilization). In our case, this analysis is useful to find out what nonverbal and verbal cues are actually capturing such kind of information. To help the analysis, we also measure the level of cue utilization with the number of significant effects.

Audiovisual nonverbal cues

Table 4.7 presents the Pearson's correlations between our sets of audio, visual, and multimodal cues and the vloggers' personality impressions. We start by providing several observations with respect to cue utilization. First, we observed that considering audio, visual, and multimodal cues together, Extraversion (cue utilization = 24) and Emotional Stability (cue utilization = 3) are the traits that show the largest and the lowest cue utilization respectively. The results concur with a recurrent finding that among all traits, Extraversion and Emotional Stability have the largest and lowest amount of informative cues respectively, in a variety of contexts [Borkenau and Liebler, 1992, Kenny et al., 1992, Ambady et al., 1995].

Second, we noticed that whereas Agreeableness was found to achieve the second largest

Speaking Activity	*E*	*C*	*O*	*A*	*ES*
Speaking Time	.18***	.27***	.13*	.05	.12**
Av Len Speak	.16***	.15**	.11†	.01	.07
# Speech turns	−.13**	−.01	−.07	.03	−.00
Cue utilization	3	2	2	0	1
Prosodic cues	*E*	*C*	*O*	*A*	*ES*
Voice rate	.02	.10†	.05	.09	.04
Av Voice Seg	−.07	−.11†	−.07	−.09†	−.06
Energy (m)	.28***	−.04	.11†	−.08	.01
Energy (m-sd)	.09†	−.13†	.08	−.01	−.00
D Energy (m)	−.08	.02	−.08	.01	−.00
D Energy (m-sd)	.34***	−.08	.15**	−.11†	−.04
Energy (max)	.32***	−.13†	.13**	−.14**	−.08
Energy (min)	.02	.01	.01	−.05	−.03
Spec Entropy (m)	.09	−.08	−.02	−.04	−.01
Spec Entropy (m-sd)	.05	.01	.06	.05	−.03
F0 (m)	.21***	−.08	.07	.12†	−.05
F0 (m-sd)	−.10†	.01	−.02	−.18**	.00
F0 (max)	.20***	.04	.14	−.00	.06
F0 (min)	.21***	−.02	.07	.01	−.02
F0 conf (m)	.22***	.02	.07	.13*	.02
F0 conf (sd)	.17***	.01	.08	.13*	.03
F0 BW (m)	−.26***	−.13*	−.15*	.06	−.02
F0 BW (m-sd)	.04	−.00	.01	.09	.03
F0 Intensity (m)	.14*	.12†	.09	−.00	.07
F0 Intensity (m-sd)	.21***	.07	.14*	−.02	.04
# R0 peaks (m)	.17***	−.09	−.01	−.04	−.04
# R0 pks (s)	.09	−.12*	−.05	−.08	−.04
Loc R0 pks (m)	.28***	−.07	.06	.03	−.07
Loc R0 pks (m-sd)	−.04	−.14*	−.07	−.14*	−.06
Cue utilization	14	8	5	8	0
Looking & Pose	*E*	*C*	*O*	*A*	*ES*
Looking time	−.02	.24***	−.03	.10†	.09
Av Len Look Seg	−.13*	.23***	−.14*	.07	.07
Proximity to camera	−.02	.07	−.05	.01	−.05
Vertical framing	.14*	.00	.14*	.12†	.08
Cue utilization	2	2	2	2	0
Visual activity	*E*	*C*	*O*	*A*	*ES*
wMEI (e)	.33***	−.17**	.21***	−.01	−.03
wMEI (m)	.32***	−.13**	.24***	.02	−.00
wMEI H Cog	.05	−.04	−.01	−.06	.01
wMEI V Cog	−.04	−.05	−.08	−.03	−.06
Cue utilization	2	2	2	0	0
Multimodal cues	*E*	*C*	*O*	*A*	*ES*
L&S	.14**	.29***	.05	.08	.12*
L&NS	−.16**	−.05	−.11	.06	−.07
L&S/L&NS	.21***	.20***	.14†	−.02	.12†
Cue utilization	3	2	1	0	2

Table 4.7: Pearson's correlation coefficients between audiovisual nonverbal cues and the personality impressions ($^{\dagger}p < .05$,* $p < .01$,** $p < .001$,*** $p < .0001$); m = mean, s = standard deviation, m-sd = mean-scaled standard deviation. For feature definition, please refer to Section 4.4.1.

agreement (ICC(1,k) = .64), it only accounted for a small number of significant effects (cue utilization = 10) compared to Conscientiousness (ICC(1,k) = .45, cue utilization = 16) and Openness to Experience (ICC(1,1) = .47 cue utilization = 12). This finding suggests that despite the evidence that certain aspects of vloggers' behavior lead annotators to agree on their impression of Agreeableness, such information is not captured by the features extracted in our data. Instead, it seems that the audiovisual cues provides more information related to impressions of Conscientiousness and Openness to Experience. More detailed observations can be found by looking at the different sets of nonverbal cues.

Among audio cues, the speaking activity features show significant correlations with all the personality impressions except with Agreeableness. This result concurs with a consistent finding in social psychology research that speaking activity cues are related to many different social constructs in conversational scenarios [Knapp and Hall, 2005]. In particular, the positive correlation of speaking time with Extraversion indicates that observers are aware of the general knowledge that associates extraverts with being talkative [Knapp and Hall, 2005]. Other correlations are also backed up by related literature. For example, Extraversion judgments in Table 4.7 are positively correlated with the length of speech segments and negatively correlated with the number of speaking turns, which agrees with findings that associate Extraversion impressions with high fluency [Scherer, 1979].

Some of the effects observed for prosodic cues had also been previously documented in the literature for other settings. The positive correlation between Extraversion and both mean/max Energy and mean Pitch is related to Extraversion impressions being associated with people speaking louder [Borkenau and Liebler, 1992, Scherer, 1979] and with higher pitch [Scherer, 1979]. The positive correlation between Agreeableness and mean Pitch is related to Agreeableness impressions associated with high voice [Borkenau and Liebler, 1992]. In addition, the positive correlation between Extraversion and the mean-scale standard deviations of Energy is associated with the idea that extraverts have higher vocal control [Knapp and Hall, 2005].

The looking patterns shown in Table 4.7 present significant correlations with all the traits except Emotional Stability. In particular, Conscientiousness impressions are associated with vloggers facing the camera longer and more persistently, as measured by the total looking time and the length of looking segments. On the contrary, Extraversion and Openness to Experience judgments are negatively correlated with the length of looking segments. One could argue that the length of looking segments is indicative of gaze avoidance. If that was the case, our results would suggest that Extraverted and Openness to Experience impressions of vloggers are associated with camera avoidance. We found at least one work showing that the association with camera avoidance was negative for most of the personality impressions [Borkenau and Liebler, 1992]. One would expect this to be true also for vlogging, specially when people are voluntarily recording themselves. However, it is unclear to what extent the distribution of this nonverbal cue may not be due to body movement or any other behavior instead of gaze avoidance. Clearly, this result needs to be explored in further detail. Regarding pose, the positive correlation of Extraversion, Openness to Experience, and Agreeableness with the

vertical framing cue suggests that high scores on these traits are associated with vloggers showing the upper body, as opposed to mainly showing the face. Interestingly, research in video-based dyadic conversations reported upper body framing to have a significant effect on participants' empathy during interaction [Nguyen and Canny, 2009].

The visual activity cues shown in Table 4.7 are among all visual cues, the ones showing the highest correlation values, doing so with Extraversion, as well as with Openness to Experience, and Conscientiousness impressions. As measured by the mean and entropy of wMEI features, high scores of Extraversion and Openness to Experience impressions are associated with high and more diverse visual activity, whereas high scores on Conscientiousness impressions are associated with a vlogging setting involving less and less diverse movement. Apparently, observers seem to share the common knowledge that associates higher levels of activity with enthusiastic, energetic, and dominant people [Knapp and Hall, 2005]. The same exact findings are reported in related literature, which show that rapid body movements are positively correlated to Extraversion impressions and negatively correlated with Conscientiousness [Kenny et al., 1992, Borkenau and Liebler, 1992].

The multimodal cues also showed a number of significant effects (see Table 4.7). Large amounts of looking-while-speaking time (L&S) are associated with high scores of Emotional Stability, Extraversion, and Conscientiousness, whereas large amounts of looking-while-not-speaking (L&NS) are also associated with low Extraversion. Note that the total looking time as a single feature (Table 4.7) did not show any correlation with Extraversion impressions but it does so when combined with speech. Furthermore, the ratio of L&S/L&NS is the multimodal cue showing the largest number of significant correlations, doing so with all the personality impressions, except Agreeableness. In particular, the results regarding Extraversion agree with previous findings that associate Extraversion impressions with people looking more frequently and with larger glances when speaking (high L&S) [Iizuka, 1992]. In addition, they concur with findings linking Extraversion to dominant behaviors. For example, in conversational scenarios, higher ratios between looking-while-speaking and looking-while-listening have been found to be associated to impressions of dominance [Dovidio and Ellyson, 1982a].

Summing up, our analysis shows that a number of automatically computed audio, visual, and multimodal nonverbal cues are significantly correlated with vloggers' personality impressions, suggesting that the behaviors measured by these cues may have also been used by the observers. As in related literature, we found that Extraversion impressions showed a significant number of associations to cues from both audio [Scherer, 1979, Borkenau and Liebler, 1992] and video [Borkenau and Liebler, 1992], whereas Conscientiousness, Agreeableness, and Openness to Experience impressions showed more associations with visual cues [Kenny et al., 1992]. We also found that the low number of correlations of cues with Agreeableness does not explain the high agreement achieved by the annotators compared to other traits, which supports the need to look for other cues.

	THR			*HMM*		
Expression	*Med*	*SD*	Q_3	*Med*	*SD*	Q_3
Anger	.47	.25	.67	.02	.21	.06
Contempt	.88	.17	.95	.42	.41	1.00
Disgust	.33	.27	.62	.01	.14	.04
Fear	.50	.26	.68	.03	.17	.09
Joy	.54	.28	.75	.03	.24	.10
Neutral	.93	.14	.98	1.00	.23	1.00
Sad	.86	.16	.94	.20	.41	1.00
Surprise	.86	.16	.94	.20	.41	1.00
Smile	.12	.15	.26	.32	.15	.41

Table 4.8: Median, SD, and third-quartile (Q_3) of PT values obtained from THR and HMM segmentations. High values for THR indicate that facial emotion activations overlap in time, whereas activations from HMM are less frequent and overlap less.

Facial Expression Cues

Before addressing the study of correlations with nonverbal behavior, we measured some basic statistics of facial activity cues with the purpose of interpreting the features obtained from THR and HMM segmentations, but also to understand the type of facial expressions that can be typically found in the vlogging scenario.

Table 5.3 reports values of the PT (proportion of time active) for cues obtained from both THR and HMM segmentations. We note that THR and HMM segmentations provide substantially different values for most expressions, and that they to agree on the high presence of the Neutral expression: in Table 5.3, around half of the vloggers show neutral expressions between 93 and 100% of the time. For the THR segmentation, this result challenges the fact that, as seen from the median values, a large number of the vlogs also shows large presence of other facial expressions, indicating that the intervals in which facial expressions are active do overlap.

In contrast, the HMM segmentation suggests a more realistic scenario in which the activation of facial expressions signals does not concur with the activation of the neutral signal. Though median values for HMM may seem low compared to THR segmentations, the third quartile (Q_3) indicates that 25% of the vlogs still show large presence of facial expressions of emotion such as Contempt, Joy, Sadness, Surprise, and Smiles.

The correlation analysis of facial cues is split between Table 4.9 and Table 4.10. As with the audiovisual cues, the facial expressions of emotion showed mostly significant effects for Extraversion independently of the representation. This trait was the one with the largest cue utilization (STATS, cue utilization = 36, THR, cue utilization = 18, and HMM, cue utilization = 19), followed by Openness to Experience and Agreeableness.

The Extraversion trait was mostly negatively correlated with features that express Anger (Mean, −.20, thr-PT; −.16, hmm-PT, −.22), and Disgust (Mean, −.12; thr-PT, −.13; hmm-PT, −.12), and positively correlated with Joy (Mean, .19; Max, .39; thr-PT, .23; hmm-PT, .23) and Smile

FE	STATS	THR	HMM
		Extraversion	
ANGER	Mean (-0.20^{**}), Med (-0.22^{**}), Min (-0.29^{***}), Var (-0.16^{*}), En ($-0.13^{\dagger}$)	PT (-0.16^{*}), PTS (0.25^{***}), NS (-0.20^{**}), AD (0.22^{**})	PT (-0.22^{**}), AD ($-0.13^{\dagger}$)
COMTEMPT	Max (0.19^{**}), Min ($-0.14^{\dagger}$), Var ($0.12^{\dagger}$)	NS ($-0.13^{\dagger}$)	PT (-0.15^{*}), PTS (0.22^{**}), NS (0.26^{***}), AD ($-0.13^{\dagger}$)
DISGUST	Mean ($-0.12^{\dagger}$), Med ($-0.12^{\dagger}$), Min ($-0.13^{\dagger}$)	PT ($-0.13^{\dagger}$), PTS (0.29^{***}), AD (0.24^{***})	PT ($-0.12^{\dagger}$)
FEAR	Mean (0.19^{**}), Max (0.36^{***}), Var (0.24^{***}), En (0.27^{***})	PT (0.22^{**})	PTS (0.25^{***}), NS (0.31^{***})
JOY	Mean (0.19^{*}), Max (0.39^{***}), Var (0.22^{**}), En (0.30^{***})	PT (0.23^{***}), PTS ($0.14^{\dagger}$), AD (0.19^{*})	PTS (0.23^{***}), NS (0.29^{***})
NEUTRAL	Mean (-0.19^{*}), Med (-0.18^{*}), Min (-0.23^{***}), Var (0.19^{*}), En ($0.12^{\dagger}$)	NS (-0.21^{**})	PTS ($0.15^{\dagger}$), NS (0.16^{*}), AD (-0.16^{*})
SAD	Max (0.25^{***}), Var (0.22^{**}), En (0.15^{*})		PTS ($0.12^{\dagger}$), NS (0.18^{*})
SMILE	Mean (0.23^{***}), Med (0.22^{**}), Max (0.31^{***}), Var (0.26^{***}), En (0.21^{**})	PT (0.25^{***}), PTS ($0.14^{\dagger}$), AD (0.18^{*})	PT (0.23^{***}), NS (0.24^{***})
SURPRISE	Mean (0.19^{**}), Med ($0.12^{\dagger}$), Max (0.31^{***}), Var (0.20^{**}), En (0.24^{***})	PT (0.16^{*}), AD ($0.11^{\dagger}$)	PTS (0.19^{**}), NS (0.26^{***})
	Cue utilization = 37	Cue utilization = 18	Cue utilization = 20
		Conscientiousness	
ANGER	Max (-0.16^{*})		
COMTEMPT	Min (0.16^{*})	NS ($0.13^{\dagger}$)	AD ($0.14^{\dagger}$)
JOY	Min ($0.12^{\dagger}$)	PT ($0.12^{\dagger}$), PTS ($-0.13^{\dagger}$), NS ($0.13^{\dagger}$)	
SAD	Max ($-0.14^{\dagger}$), Var (-0.15^{*}), En (-0.18^{*})		NS ($-0.11^{\dagger}$)
SMILE	Min ($0.12^{\dagger}$)		
NEUTRAL		NS (0.18^{*}), AD ($-0.13^{\dagger}$)	
SURPRISE		PTS ($-0.13^{\dagger}$), AD ($-0.12^{\dagger}$)	
	Cue utilization = 7	Cue utilization = 8	Cue utilization = 2

Table 4.9: R-squared results with SVM and RF.. ($^{\dagger}p < .05,^{*} p < .01,^{**} p < .001,^{***} p < .0001$). For feature definition, please refer to Section 4.4.2.

(Mean, .23; thr-PT, .25; hmm-PT, .23) which concurs with the idea that Extraverted are more enthusiastic people. However, other effects may be more difficult to explain, such as the positive correlation with Fear (Mean, .19; thr-PT, .22), or with Sad (Max, .25; hmm-PT, .25). Openness to Experience also showed similar negative correlations for Anger (Mean, −.19, thr-PT; −.20, hmm-PT, −.16), but showed only a couple of effects of Joy and Smile. We also observed positive correlations between Openness to Experience and Fear, which concur with the effects in Extraversion and may suggests that this facial expression is not correctly estimated, though this needs further investigation.

The Agreeableness trait is also negatively correlated with Anger (Mean, −.14; PT,−.16) and

FE	STATS	THR	HMM
		Openness to Experience	
ANGER	Mean (-0.19^{**}), Med (-0.20^{**}), Min (-0.22^{**}), Var (-0.16^{*}), En (-0.18^{*})	PT (-0.20^{**}), PTS ($0.13^{\dagger}$), NS (-0.17^{*}), AD ($0.12^{\dagger}$)	PT (-0.16^{*})
DISGUST	En ($-0.13^{\dagger}$)	PT ($-0.13^{\dagger}$), PTS ($0.13^{\dagger}$), AD ($0.11^{\dagger}$)	NS ($-0.12^{\dagger}$)
FEAR	Mean ($0.13^{\dagger}$), Max (0.21^{**}), Var ($0.14^{\dagger}$), En (0.23^{***})	PT (0.19^{**})	PTS (0.15^{*}), NS (0.18^{*})
JOY	Max (0.21^{**}), En (0.20^{**})	PT (0.15^{*})	
SAD	Max (0.18^{*})		
SMILE	Var ($0.12^{\dagger}$)		PT ($0.14^{\dagger}$)
SURPRISE	Mean (0.17^{*}), Med ($0.14^{\dagger}$), Max (0.18^{*}), Var ($0.14^{\dagger}$), En (0.22^{**})	PT (0.17^{*}), NS ($0.12^{\dagger}$)	PTS (0.16^{*}), NS (0.17^{*})
	Cue utilization = 19	Cue utilization = 11	Cue utilization = 7
		Aggreableness	
ANGER	Mean ($-0.14^{\dagger}$), Max (-0.17^{*}), Var ($-0.14^{\dagger}$), En (-0.19^{**})	PT (-0.16^{*})	NS (-0.20^{**})
DISGUST	Max ($-0.14^{\dagger}$)		PTS (-0.17^{*}), NS ($-0.15^{\dagger}$)
JOY	Mean (0.20^{**}), Med (0.18^{*}), Max ($0.12^{\dagger}$), Min ($0.13^{\dagger}$), Var (0.16^{*}), En (0.21^{**})	PT (0.18^{*}), PTS ($-0.13^{\dagger}$), NS ($0.14^{\dagger}$)	PT ($0.13^{\dagger}$)
SMILE	Mean (0.18^{*}), Med (0.19^{*}), Min (0.17^{*}), En (0.20^{**})	PT (0.20^{**}), AD ($0.12^{\dagger}$)	PT ($0.12^{\dagger}$), AD (0.16^{*})
SAD			PTS ($-0.11^{\dagger}$)
	Cue utilization = 15	Cue utilization = 6	Cue utilization = 7
		Emotional Stability	
JOY	Mean (0.15^{*}), Med (0.15^{*}), Var ($0.11^{\dagger}$)	PTS ($-0.15^{\dagger}$), NS ($0.14^{\dagger}$), AD ($-0.14^{\dagger}$)	
SMILE	Min ($0.13^{\dagger}$)	PT ($0.12^{\dagger}$)	AD (0.18^{*})
ANGER			NS ($-0.15^{\dagger}$)
DISGUST			PTS (-0.17^{*}), NS ($-0.12^{\dagger}$)
SAD			PTS ($-0.14^{\dagger}$)
	Cue utilization = 4	Cue utilization = 4	Cue utilization = 5

Table 4.10: R-squared results with SVM and RF.. ($^{\dagger}p < .05, ^{*}p < .01, ^{**}p < .001, ^{***}p < .0001$). For feature definition, please refer to Section 4.4.2.

positively correlated with Joy (Mean, .20; PT,.18). and Smile (Mean, .18; En .20; thr-PT,.20), and did not show any effects with any other expressions. Finally, Conscientiousness and Emotional Stability showed a very small number of effects. Overall, we found that THR and HMM features provided similar effects in terms of the sign, though the cue utilization varied across facial expressions and trait.

In summary, whereas CERT seems to be capturing information that agrees with impressions of personality, it is unclear how much the method suffers from processing challenging conversational social video like the one we study. In particular, the fact that most people talk a substantial amount of time may trigger some facial expressions due to lip movements that would not be otherwise activated. These issues may be investigated in future work.

Trait	LIWC categories
Extr	tentat (-0.19^{***}), nonfl (-0.18^{**}), cogmech (-0.14^{*}), discrep (-0.13^{*}), excl ($-0.12^{\dagger}$), ipron ($-0.11^{\dagger}$), health ($-0.10^{\dagger}$), social ($0.10^{\dagger}$), affect ($0.11^{\dagger}$), assent (0.13^{*}), you (0.13^{*}), space (0.16^{**}), sexual (0.17^{**}) Cue utilization = 12
Cons	assent (-0.23^{***}), i (-0.23^{***}), filler (-0.22^{***}), negemo (-0.19^{***}), ppron (-0.19^{***}), negate (-0.18^{**}), verb (-0.17^{**}), anger (-0.17^{**}), pronoun (-0.15^{*}), present (-0.15^{*}), swear (-0.15^{*}), adverb (-0.14^{*}), sexual (-0.14^{*}), auxverb (-0.14^{*}), time (-0.13^{*}), body ($-0.10^{\dagger}$), they ($0.11^{\dagger}$), discrep ($0.12^{\dagger}$), article (0.12^{*}), incl (0.17^{**}), achieve (0.17^{**}), work (0.21^{***}), preps (0.24^{***}), Sixltr (0.25^{***}) Cue utilization = 24
Open	health (-0.14^{*}), anger (-0.13^{*}), negemo (-0.12^{*}), nonfl ($-0.12^{\dagger}$), sad ($-0.11^{\dagger}$), swear ($-0.10^{\dagger}$), death ($-0.10^{\dagger}$), hear ($0.10^{\dagger}$), motion ($0.10^{\dagger}$), leisure (0.13^{*}) Cue utilization = 10
Agr	anger (-0.43^{***}), negemo (-0.42^{***}), swear (-0.37^{***}), sexual (-0.28^{***}), bio (-0.17^{**}), negate (-0.14^{*}), relig (-0.14^{*}), they (-0.13^{*}), quant ($-0.11^{\dagger}$), work ($0.09^{\dagger}$), friend ($0.11^{\dagger}$), incl ($0.12^{\dagger}$), i ($0.12^{\dagger}$), conj (0.14^{*}), posemo (0.24^{***}) Cue utilization = 15
Emot	negemo (-0.38^{***}), anger (-0.34^{***}), swear (-0.31^{***}), sexual (-0.31^{***}), bio (-0.17^{**}), negate (-0.16^{**}), affect ($-0.10^{\dagger}$), nonfl ($0.09^{\dagger}$), discrep ($0.10^{\dagger}$), work ($0.12^{\dagger}$), leisure ($0.12^{\dagger}$), achieve (0.12^{*}) Cue utilization = 12

Table 4.11: Selection of significant Pearson's correlation effects (p <.05) between LIWC cues personality impressions.

Verbal Content Cues

Table 4.11 summarizes the significant correlations ($p < .05$) between LIWC categories and Big Five scores (from most negative to most positive). We did not look at correlations with unigrams and bigrams because of the sparsity of these features. We found a total of 12 significant effects for judgments of Extraversion, 8 out of which are backed up by the literature. For example, we found that Extraversion judgments were associated with an increased use of categories related to interpersonal interaction (*you*, $r = .13$, *social*: $r = .10$) [Pennebaker and King, 1999, Gill et al., 2009a]. As in [Gill et al., 2009a], we found that Extraversion is the only trait associated with the use of 2nd person of singular (i.e., vloggers refer frequently to the YouTube audience). The increased use of sexual words (*sexual*, $r = .17$,) associated to the Extraversion trait is also documented in previous work [Yarkoni, 2010]. We also found that vloggers judged as introverted use more cognitive related words (*cogmech*: $r = -.14$) , including discrepancy (*discrep*: $r = -.13$), tentative (*tentat*: $r = -.19$), and exclusive words (*excl*: $r = -.12$), concurring with previous literature [Yarkoni, 2010]. In addition, as in face-to-face interactions [Mehl et al., 2006], we found Extraversion judgments associated with the expression of emotions (*affect*: $r = .11$).

Conscientiousness judgments showed 24 significant effects. Not surprisingly, we found Conscientiousness judgments to show some of the largest associations with words related to

occupation and achievement (*work*: $r = .21$, *achieve*: $r = .17$) which is consistent with findings that associate Conscientiousness to an increase usage of these word categories [Gill et al., 2009a]. These vloggers are also associated to a decreased use of negative emotion words (*negate*: $r = -.18$, *negemo*: $r = -.19$) [Mehl et al., 2006], swearing words (*swear*: $r = -.15$), and sexual words (*sexual*: $r = -.14$) [Yarkoni, 2010]. Though we did not find any positive association between Conscientiousness and the 3rd person pronoun as in [Gill et al., 2009a], we found a negative association on the use of the 1st person pronoun (*i*: $r = -.23$). Other effects, not documented in the literature, are the correlation with auxiliary verbs (*auxverb*: $r = -.14$) and present tense (*present*: $r = -.15$), and the positive association with prepositions (*preps*: $r = .24$), articles (*article*: $r = .12$), and inclusive (*incl*: $r = .17$). We also found that this trait was the only one positively correlated with the length of the words (*Sixltr*, $r = .25$), which [Mehl et al., 2006] associate to a careful choice of words.

We found 10 effects for Openness to Experience judgments. Vloggers judged as opened to experience tend to use more words related to topics focused on leisure activities (*leisure*: $r = .13$) [Gill et al., 2009a], and words concerning the senses (hear : $r = .13$) [Gill et al., 2009a]. In addition, they tend to express negative emotions less frequently (*anger*: $r = -.13$, *negemo*: $r = -.15$, *anger*: $r = -.13$) [Pennebaker and King, 1999, Gill et al., 2009a].

Agreeableness judgments displayed 15 significant effects with LIWC (including categories and sub-categories). First, as shown in previous literature [Pennebaker and King, 1999], we found the largest effects for Agreableness judgments with the use of both positive (*posemo*: $r = .24$), and negative emotions (*anger*: $r = -.43$, *negemo*: $r = -.42$). This relates to the idea that agreeable people are socially oriented and tend to avoid conflict [Mehl et al., 2006]. Indeed, concurring with previous work, we also found positive associations with the use of self references (*i*: $r = .12$) [Yarkoni, 2010], friendship (*friend*: $r = .11$) [Mehl et al., 2006], and a negative association with *they* ($r = -.13$). In contrast, the large correlations with *anger* ($r = -.43$) and *negemo* ($r = -.42$) categories show that annotators associated the use of these type of words with more disagreeable people. In addition, disagreeable people also use more words related to sexuality (*sexual*: $r = -.29$), swear words (*swear*: $r = -.37$), body states (*bio*: $r = -.17$), and religion (*relig*: $r = -.14$).

Finally, Emotional Stability showed 12 significant effects. High emotional scorers of this trait are associated to the expression of negative words (*negate*: $-.16$), negative emotional words (*anger*: $r = -.34$, *negemo*: $r = -.30$, *affect*: $r = -.10$) [Pennebaker and King, 1999], swear words (*swear*: $-.31$) and sexual works (*sexual*: $-.31$).

Overall, our work backs up findings that relate nonverbal behavior and verbal content to personality impressions in works from social media and social psychology. We noted that the magnitude of the observed effects in nonverbal behavior compare modestly (yet stastitical significant) with effects reported in some social psychology works [Ambady et al., 1995, Scherer, 1979, Kenny et al., 1992, Borkenau and Liebler, 1992], which report significant correlations between .15 and .70 for a diversity of nonverbal cues. This could be due to different factors.

For example, as observed by Yarkoni [Yarkoni, 2010], this could be explained by the fact that effect sizes for statistical significant effects typically vary inversely with the sample size. For example, [Borkenau and Liebler, 1992] investigated the personality correlates on a sample of N=100 people and most significant effects ($p < .05$) had correlation values between 20 and 50. In our case, the highest correlations for nonverbal cues were $r = .33$ and $r = .39$ for audiovisual and facial cues respectively. However, it could also be partially explained by the fact that the cues automatically extracted are fine-grained when computed from audiovisual analysis, whereas in most social psychology works, the constructs were assessed using manual scales. The effects found analyzing verbal content from vlogs compare to verbal content analysis from text blogs in [Yarkoni, 2010], which might support this hypothesis.

4.6.3 Automatic Prediction of Personality Impressions

This section is divided in four parts. First, we present results on automatically predicting personality impressions using nonverbal cues and verbal content cues independently. Then, by focusing on the subsample of vloggers included in all feature sets, we compare the performance of modalities and investigate possible effects when combining them.

In all tables, R^2 values are averaged over the 10 test folds, and values between parenthesis correspond to the standard deviation computed on the 10 fold means. To measure significant differences between the models and the baseline, we conducted two-tailed paired t-tests for the RMSE, and two-tailed single t-tests for R^2, and we include p-values in the tables.

Using Audiovisual Nonverbal Cues

Table 4.12 summarizes the performance on the prediction of the Big-Five personality impressions using different sets of audiovisual nonverbal cues (N=442). RFs provided better results than SVMs by a very small margin.

At a first glance, the best performance (up to $R^2 = 39\%$) was achieved for Extraversion, which is not surprising, given that it is the trait with the largest cue utilization, and the one that achieved the most agreement among observers. The performance degrades for the second best predicted traits, Conscientiousness and Openness to Experience, that achieve up to $R^2 = 10\%$. Finally, although statistically significant, the performance for Agreeableness and Emotional Stability is low with respect to the baseline.

Among all audiovisual feature sets, prosodic cues (PR) are the best performing single features, achieving $R^2 = 33\%$ for Extraversion. In comparison, the best results for Conscientiousness were obtained with speaking activity cues (SA) and SVMs ($R^2 = 10\%$). Regarding visual cues, visual activity (VA) was useful for the prediction of the Extraversion trait (up to $R^2 = 12\%$) and very slightly for Openness to Experience (up to $R^2 = 5\%$), whereas Look and Pose cues (LP) provided poor performance in general. Also for Extraversion, the performance of multimodal (M) cues was at least as good as using speaking activity (SA) or looking cues in combination

Features	Extr		Cons		Open		Agr		Emot	
Baseline	.00		.00		.00		.00		.00	
SVMRadial										
SA	$.05^{\dagger}$	(.09)	$.10^{**}$	(.06)	$.02^{\dagger}$	(.03)	$.02^{\dagger}$	(.02)	$.06^{*}$	(.05)
PR	$.32^{***}$	(.06)	$.07^{**}$	(.03)	$.06^{*}$	(.04)	$.05^{\dagger}$	(.05)	$.00^{\dagger}$	(.01)
A (SA+PR)	$.33^{***}$	(.06)	$.07^{**}$	(.04)	$.06^{*}$	(.04)	$.05^{\dagger}$	(.05)	$.01^{\dagger}$	(.01)
LP	$.05^{\dagger}$	(.06)	$.04^{**}$	(.02)	$.03^{\dagger}$	(.04)	$.03^{*}$	(.03)	$.01^{\dagger}$	(.02)
VA	$.12^{**}$	(.08)	$.01^{\dagger}$	(.01)	$.05^{*}$	(.05)	$.01^{\dagger}$	(.02)	$.01^{\dagger}$	(.01)
V (LP+VA)	$.10^{***}$	(.04)	$.05^{*}$	(.04)	$.06^{*}$	(.04)	$.03^{\dagger}$	(.03)	$.01^{\dagger}$	(.01)
M	$.07^{***}$	(.02)	$.09^{**}$	(.05)	$.03^{\dagger}$	(.04)	$.02^{*}$	(.02)	$.03^{\dagger}$	(.04)
A+V	$.37^{***}$	(.06)	$.09^{***}$	(.03)	$.10^{**}$	(.06)	$.06^{\dagger}$	(.06)	$.01^{*}$	(.01)
A+M	$.33^{***}$	(.06)	$.09^{**}$	(.04)	$.06^{**}$	(.04)	$.05^{\dagger}$	(.05)	$.02^{\dagger}$	(.02)
V+M	$.17^{**}$	(.09)	$.07^{*}$	(.05)	$.07^{**}$	(.04)	$.03^{\dagger}$	(.04)	$.02^{\dagger}$	(.02)
A+V+M	$.36^{***}$	(.05)	$.10^{***}$	(.04)	$.10^{**}$	(.06)	$.06^{\dagger}$	(.06)	$.02^{*}$	(.01)
RF										
SA	$.04^{\dagger}$	(.04)	$.04^{\dagger}$	(.04)	$.02^{\dagger}$	(.03)	$.01^{\dagger}$	(.02)	$.02^{\dagger}$	(.04)
PR	$.34^{***}$	(.07)	$.07^{*}$	(.06)	$.07^{*}$	(.07)	$.05^{\dagger}$	(.05)	$.01^{\dagger}$	(.02)
A (SA+PR)	$.34^{***}$	(.07)	$.09^{*}$	(.07)	$.07^{\dagger}$	(.08)	$.06^{\dagger}$	(.06)	$.01^{\dagger}$	(.01)
LP	$.04^{\dagger}$	(.05)	$.00^{\dagger}$	(.01)	$.03^{\dagger}$	(.04)	$.02^{\dagger}$	(.02)	$.01^{\dagger}$	(.01)
VA	$.11^{\dagger}$	(.10)	$.02^{*}$	(.02)	$.05^{*}$	(.04)	$.01^{*}$	(.01)	$.01^{*}$	(.01)
V (LP+VA)	$.12^{*}$	(.08)	$.02^{\dagger}$	(.02)	$.05^{**}$	(.03)	$.01^{\dagger}$	(.01)	$.02^{\dagger}$	(.04)
M	$.06^{**}$	(.03)	$.05^{*}$	(.04)	$.03^{\dagger}$	(.04)	$.01^{\dagger}$	(.02)	$.02^{\dagger}$	(.02)
A+V	$.39^{***}$	(.08)	$.10^{**}$	(.05)	$.10^{*}$	(.07)	$.04^{*}$	(.04)	$.01^{\dagger}$	(.01)
A+M	$.35^{***}$	(.06)	$.09^{*}$	(.06)	$.07^{\dagger}$	(.07)	$.05^{*}$	(.04)	$.01^{\dagger}$	(.01)
V+M	$.20^{**}$	(.12)	$.04^{*}$	(.03)	$.06^{**}$	(.04)	$.02^{\dagger}$	(.03)	$.02^{\dagger}$	(.02)
A+V+M	$.38^{***}$	(.08)	$.09^{**}$	(.05)	$.09^{*}$	(.06)	$.05^{\dagger}$	(.05)	$.01^{\dagger}$	(.02)
Highest achieved performance										
	.39		.10		.10		.06		.06	

Table 4.12: R-squared results on predicting personality impressions using SVM and RF for audiovisual nonverbal cues (SA = Speaking Activity, PR = Prosody, LP = Look and pose, VA = Visual Activity, M = Multimodal), $^{\dagger}p < .05$,* $p < .01$,** $p < .001$,*** $p < .0001$.

with pose (LP). Finally, the combination of audio, visual, and multimodal cues showed only small improvements with respect to the use of single feature sets alone. The largest improvement with respect to using a single feature set, was obtained using visual and multimodal cues (V+M, up to $R^2 = 20\%$ for Extraversion), while the highest performance was achieved when combining audio and visual cues (A+V, $R^2 = 39\%$ for Extraversion). Adding multimodal cues to any combination including audio did not result in any improvement.

4.6.4 Using Facial Expressions Cues

Table 4.13 summarizes the performance on predicting personality impressions using cues derived from facial expressions of emotion, smile, and their combination (N=298). First, we

Features	Extr		Cons		Open		Agr		Emot	
Base	.00		.00		.00		.00		.00	
				SVMRadial						
STATS-FE	.22**	(.12)	.04†	(.04)	.11†	(.11)	.06†	(.11)	.07†	(.07)
THR-FE	.19*	(.14)	.06†	(.08)	.08†	(.08)	.06†	(.08)	.04*	(.03)
HMM-FE	.16*	(.14)	.03*	(.03)	.08**	(.05)	.07†	(.09)	.04*	(.03)
STATS-smile	.10*	(.08)	.02†	(.02)	.02†	(.02)	.06†	(.07)	.03†	(.04)
THR-smile	.09†	(.10)	.05*	(.04)	.06*	(.06)	.05*	(.04)	.04†	(.04)
HMM-smile	.06†	(.07)	.03†	(.05)	.08*	(.07)	.05†	(.07)	.07†	(.11)
STATS	.23**	(.14)	.04†	(.05)	.11†	(.11)	.08†	(.08)	.06*	(.06)
THR	.22*	(.17)	.06†	(.08)	.08†	(.08)	.08†	(.08)	.04*	(.03)
HMM	.17*	(.13)	.03*	(.02)	.07*	(.06)	.07†	(.09)	.03†	(.04)
HMM+thr	.23*	(.18)	.04†	(.05)	.09*	(.09)	.06†	(.08)	.05†	(.08)
STATS thr hmm	.24*	(.18)	.05†	(.08)	.10†	(.11)	.07†	(.08)	.06†	(.07)
				RF						
STATS-FE	.23**	(.13)	.04†	(.04)	.08*	(.08)	.05†	(.05)	.02†	(.05)
THR-FE	.17*	(.12)	.06†	(.08)	.08†	(.10)	.06†	(.10)	.00†	(.01)
HMM-FE	.16*	(.12)	.02*	(.02)	.07*	(.06)	.06†	(.10)	.04*	(.02)
STATS-smile	.13*	(.10)	.02*	(.02)	.02†	(.02)	.06**	(.03)	.01†	(.01)
THR-smile	.07†	(.10)	.03†	(.04)	.05†	(.07)	.08†	(.13)	.03†	(.04)
hmm-smile	.06†	(.06)	.01†	(.01)	.04†	(.05)	.04†	(.06)	.03†	(.05)
STATS	.24**	(.13)	.02†	(.02)	.08†	(.09)	.06†	(.06)	.02†	(.03)
THR	.22*	(.16)	.04*	(.04)	.07†	(.09)	.06†	(.08)	.01†	(.01)
HMM	.18*	(.14)	.03*	(.02)	.07*	(.05)	.08†	(.10)	.06†	(.05)
HMM+THR	.23*	(.18)	.03†	(.04)	.08*	(.08)	.07†	(.09)	.04†	(.04)
STATS+THR+HMM	.23*	(.17)	.03*	(.03)	.09*	(.08)	.06†	(.07)	.03†	(.05)
				Highest achieved performance						
	.24		.06		.11		.08		.07	

Table 4.13: R-squared results on predicting personality impressions using SVM and RF for facial expression cues (FE) and smile, $^{\dagger}p < .05$,* $p < .01$,** $p < .001$,*** $p < .0001$.

noted that compared to the audiovisual nonverbal cues, Extraversion is still the trait for which we achieve the highest prediction Extraversion (R^2= 24%). This is not surprising, because the correlation analysis also show the largest cue utilization for this trait. Openness to Experience (R^2= 11%) and Agreeableness (R^2= 8%) were the second and third best predicted traits in terms of R-squared, though with substantially lower performance. For the case of Agreeableness, this is notable because this trait could not be predicted with audiovisual nonverbal cues.

We also noted that the standard deviation of these predictions was larger compared to results using the previous audiovisual nonverbal cues, which decreases the power of the statistical significance tests. We hypothesize that this may be due to the evidence discussed in the correlation analysis that these features may be more noisy than the other cues. Finally, SVMs and RFs provided similar performance.

Features	Extr		Cons		Open		Agr		Emot	
Baseline	.00		.00		.00		.00		.00	
					$SVMRadial$					
LIWC	$.14^{*}$	(.12)	$.19^{**}$	(.11)	$.04^{\dagger}$	(.04)	$.26^{**}$	(.13)	$.08^{*}$	(.07)
Unigrams	$.05^{\dagger}$	(.05)	$.14^{**}$	(.08)	$.03^{\dagger}$	(.06)	$.14^{*}$	(.09)	$.07^{*}$	(.05)
Bigrams	$.04^{\dagger}$	(.05)	$.08^{*}$	(.06)	$.03^{*}$	(.02)	$.13^{*}$	(.12)	$.04^{\dagger}$	(.04)
					RF					
LIWC	$.13^{*}$	(.10)	$.18^{**}$	(.10)	$.04^{**}$	(.02)	$.31^{***}$	(.12)	$.17^{*}$	(.13)
Unigrams	$.11^{***}$	(.04)	$.14^{***}$	(.05)	$.03^{\dagger}$	(.03)	$.21^{**}$	(.13)	$.12^{*}$	(.11)
Bigrams	$.04^{\dagger}$	(.04)	$.12^{**}$	(.06)	$.02^{\dagger}$	(.04)	$.14^{*}$	(.11)	$.11^{\dagger}$	(.10)
					Highest achieved performance					
	.14		.19		.04		.31		.17	

Table 4.14: R-squared results on predicting personality impressions using SVM and RF for LIWC, unigram, and bigram cues computed in manual speech transcriptions, $^{\dagger}p < .05$,* $p < .01$,** $p < .001$,*** $p < .0001$.

Interestingly, the set of cues proposed from THR and HMM segmentations provided inferior performance to a simple model using statistical descriptors of facial expressions and smile, which suggests that despite being simple to interpret they do not capture the potential of facial expressions of emotion. Despite the comparable cue utilization achieved by smile features with Extraversion and Agreeableness traits, the results also show that the smile feature alone is more useful to predict the Extraversion trait (up to $R^2 = 13\%$). We also found that combining smile with facial expressions improve results very little, which may indicate a large correlation between these feature and facial expressions.

Using verbal content

Table 4.14 summarizes the performance on the prediction of personality impressions using verbal content from vlogs from manual transcriptions (N=408). At a glance, the results differ from previous works in at least two aspects. The first one is that all the Big-Five personality impressions can be predicted substantially better than the baseline. The second is that instead of Extraversion, Agreeableness is the trait that shows higher performance ($R^2 = 31\%$). The result is relevant because, despite the fact that Agreeableness is the second trait with highest agreement, the audiovisual models did not predict this trait, and facial expression cues achieved a modest $R^2 = 8\%$.

In addition, the verbal content is useful to predict other traits such as Conscientiousness ($R^2 = .19$), and Emotional Stability ($R^2 = .17$), which are the best predictions for these traits. In this case, the RFs provided substantially higher performance for Agreeableness and Emotional Stability than using SVMs. Finally, though Extraversion also was predicted significantly better than the baseline, results are poor compared to other Big-Five given the differences in

Features	Extr		Cons		Open		Agr		Emot	
Baseline	.00		.00		.00		.00		.00	
					$SVMRadial$					
Uni-inCFS	$.04^{*}$	(.03)	$.06^{*}$	(.06)	$.00^{\dagger}$	(.00)	$.10^{**}$	(.06)	$.02^{\dagger}$	(.02)
Bi-inCFS	$.02^{\dagger}$	(.03)	$.01^{\dagger}$	(.01)	$.00^{\dagger}$	(.00)	$.05^{*}$	(.04)	$.04^{\dagger}$	(.04)
Uni-outCFS	$.15^{***}$	(.06)	$.12^{*}$	(.08)	$.06^{**}$	(.03)	$.18^{**}$	(.10)	$.08^{*}$	(.06)
Bi-outCFS	$.24^{**}$	(.12)	$.14^{**}$	(.08)	$.16^{***}$	(.07)	$.17^{**}$	(.09)	$.13^{*}$	(.10)
					RF					
Uni-inCFS	$.07^{*}$	(.05)	$.10^{*}$	(.09)	$.01^{*}$	(.01)	$.21^{***}$	(.09)	$.07^{\dagger}$	(.07)
Bi-inCFS	$.03^{\dagger}$	(.03)	$.04^{*}$	(.03)	$.00^{\dagger}$	(.01)	$.08^{*}$	(.05)	$.11^{\dagger}$	(.12)
Uni-outCFS	$.28^{***}$	(.06)	$.31^{***}$	(.11)	$.16^{**}$	(.09)	$.37^{***}$	(.09)	$.26^{**}$	(.17)
Bi-outCFS	$.33^{***}$	(.13)	$.32^{**}$	(.16)	$.28^{***}$	(.08)	$.41^{***}$	(.14)	$.32^{**}$	(.16)
Highest achieved performance										
	.33		.32		.28		.41		.32	

Table 4.15: R-squared results on predicting personality impressions using SVM and RF using unigrams (uni) and bigram (bi) with CFS inside cross-validation (inCFS) and outside (outCFS). Improvements for outCFS suggest overfitting. $^{\dagger}p < .05,^{*}p < .01,^{**}p < .001,^{***}p < .0001$.

Features	Extr		Cons		Open		Agr		Emot	
Base	.00		.00		.00		.00		.00	
					RF					
liwc	$.02^{**}$	(.02)	$.08^{**}$	(.06)	$.02^{*}$	(.02)	$.10^{**}$	(.08)	$.05^{**}$	(.04)
liwc-lowWER	$.04^{*}$	(.07)	$.10^{*}$	(.11)	$.05^{*}$	(.05)	$.18^{**}$	(.12)	$.10^{*}$	(.12)
liwc-highWER	$.07^{*}$	(.07)	$.10^{**}$	(.08)	$.01^{*}$	(.02)	$.12^{*}$	(.14)	$.05^{**}$	(.04)
Highest achieved performance										
	.07		.10		.05		.18		.10	

Table 4.16: R-squared results on predicting personality impressions using and RF and LIWC for automatic transcriptions. liwc-lowWER and liwc-highWER are models retrained on low WER and high WER, respectively, $^{\dagger}p < .05,^{*}p < .01,^{**}p < .001,^{***}p < .0001$.

judgment reliability, but also because the nonverbal cues seem more useful to predict this trait.

Our results also showed superior performance of LIWC compared to unigrams or bigrams. In particular, our experiments using CFS indicate that using a feature selection procedure outside the training loop as suggested by [Iacobelli et al., 2011] may result in overfitting issues. This is shown in Table 4.15, where we compared unigrams and bigram models with the use of CFS inside (inCFS) and outside (outCFS) the cross-validation procedure. The improvement is larger for the case of the bigram representation, which tends to be sparser than unigrams, and therefore more prone to larger overfitting for small amounts of data. As a result, we cannot trust the ourCFS procedure as being correct.

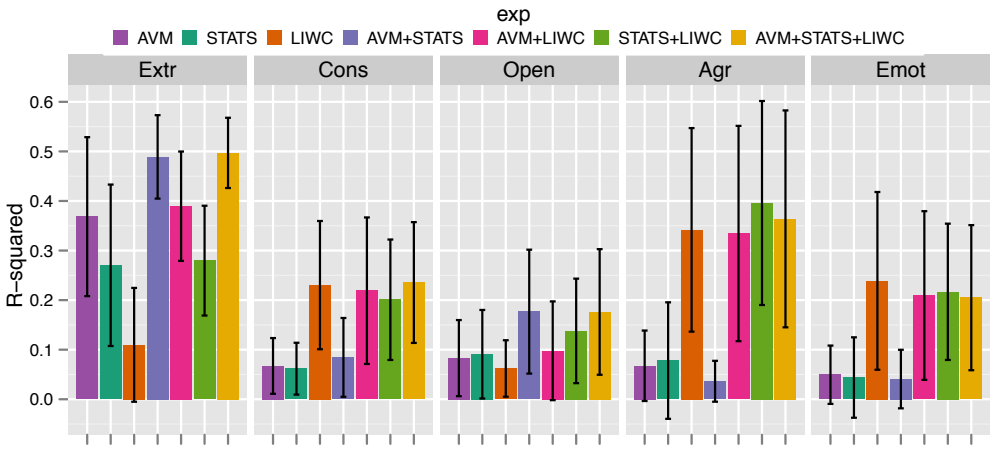

Figure 4.5: R-squared results on predicting personality impressions using RFs, best models for each modality (AVM for audiovisual, STATS for facial cues, and LIWC for verbal content), and combinations of them.

Finally, Table 4.16 shows the results when using automatic transcriptions instead of manual ones computed using LIWC and RF only, which was the best setting in previous experiments (N=397). We see that results dropped significantly when using automatic transcripts as a result of the errors introduced by the ASR system. In particular, we see a drop in performance from $R^2 = .31$ to .10, for Agreeableness, and from $R^2 = .18$ to .08 for Conscientiousness, whereas for the rest of the traits the performance is close to the baseline. The effect of the WER is clearer when retraining two models for samples with high and low WER. Compared to using all the data, the performance of the Agreeableness trait doubles for the subset with low WER, despite the fact that the WER was still high ($WER = 77\%$). The results for Emotional Stability also improve for samples with lower WER, but for the rest of the traits it remains the same.

Comparison and Fusion

Figure 4.5 shows the performance of the modalities (best individual models) when applied alone to predict personality impressions on the subsample of vloggers included within the three feature sets ($N = 295$), i.e., not considering the automatic transcriptions.

First, we note that the performances in this subset of data are marginally higher than the that results provided in the previous subsections and feature sets. Best performance for audiovisual (AVM) and facial expressions cues (STATS) was achieved for Extraversion with $R^2 = .36$ ($p < 10^{-4}$) and $R^2 = .27$ ($p < 10^{-3}$) respectively, whereas for verbal content (LIWC), best performance was achieved for Agreeableness ($R^2 = .34$, $p < 10^{-3}$), Emotional Stability ($R^2 = .24$, $p < 10^{-3}$), and Conscientiousness ($R^2 = .23$, $p < 10^{-3}$). Symbols indicated p-values corresp Overall, Extraversion performance show less variance than Agreeableness and Openness to Experience, which may be due to the combination of having higher reliability impressions and using more robust cues for prediction.

We also see that two combinations of modalities help substantially to improve the performance for Extraversion, Agreeableness, and Openness to Experience. Combining audiovisual features and facial expressions (AVM+STATS) boosts the performance of Extraversion predictions up to $R^2 = .48$ ($p < 10^{-4}$), while for the Openness to Experience, it doubles the performance of any single best predictor up to $R^2 = 17$ ($p < 10^{-2}$). In both cases, adding verbal content contributes to marginal improvements. In contrast, for Agreeableness, the performance improved when combining facial expression cues and verbal features (STATS+LIWC) from $R = .34$ ($p < 10^{-3}$) to $R = .39$ ($p < 10^{-3}$), whereas including audiovisual cues did not improve performance much.

Overall, our experiments show that the task of automatically predicting personality impressions in vlogging is feasible and that the performance of the models varies across traits and modalities, and that modality fusion can be advantageous. Our results concur with related literature on the prediction of personality impressions in several aspects (see review in Chapter 2). First, we found that Extraversion is trait predicted with highest performance [Mairesse et al., 2007, Lepri et al., 2009, Mohammadi et al., 2010]. Second, we found that prosodic cues are the best set among audiovisual nonverbal cues [Mairesse et al., 2007]. And third, that verbal content was useful to predict traits that could not be predicted otherwise with audiovisual nonverbal cues [Mairesse et al., 2007]. The achieved performances also compare well with existing attempts to predict personality. For example, Mairesse et al. [2007] obtained the best performances for the prediction of Extraversion, Openness, and Emotional Stability with R^2 values of 24%, 18% and 15% respectively based on verbal text content. In [Lepri et al., 2009], Lepri et al. obtained up to R^2 of 22% for self-reported Extraversion based on automatic nonverbal cues.

4.7 Conclusions

In this chapter we investigated the problem of interpersonal perception in vlogging with a focus on personality research and automatic behavioral analysis. First, personality impressions were leveraged to revisit the problem of social attention and inspect to what extent vlogger personality traits can explain the links between behavioral cues and social attention found in the previous chapter. Second, we investigated the automatic analysis of vlogs to extract features from nonverbal behavior and verbal content and to build computational models of personality. The rest of this section summarizes the main contributions of this chapter and discuss some limitations.

Our study revealed how personality impressions are connected to the YouTube vlog watching experience in such a way that certain vlogger traits result on audiences watching, commenting, rating, and favoring their videos more. Our results showed **positive linear associations for Extraversion, Openness to Experience, and Conscientiousness with respect to these measures, which indicate that people scoring high on these traits are found to be more interesting by audiences**. These results concur with interpersonal perception theory on the type

of personal traits that relate to people achieving attention or reacting to it in face-to-face interactions. In contrast, **the Agreeableness trait showed a U-shape relation with attention, indicating that vloggers on both extremes of these scales receive comparable treatment from audiences in terms of attention**. Future work could study the phenomenon in more detail. For example, it is interesting to see that most of the Big-Five traits are positively correlated with attention. The fact that these impressions are positively correlated among them, poses the open question of whether a first factor in personality impressions would essentially represent overall perceived positivity and thus, result in a measure of social attention.

Note that the correlation analysis between impressions and social attention resulted in larger effects than the correlation values obtained between behavioral cues and impressions. This happens because in the case of social attention we use an aggregate measure of attention rather than the individual measures of attention of videos, which results in a completely different methodology than the one used for behavioral cues.

Second, we investigated the cue utilization of different modalities with respect to crowdsourced personality impressions of vloggers. Neither the audiovisual nor the facial emotion expression aspect of users had been addressed before in the context of personality impressions and social media research. Our correlation analysis showed that **audiovisual nonverbal cues and simple statistical aggregates of facial emotion expressions are useful to explain personality impressions of Extraversion mainly**, and showed weaker associations with Conscientiousness (for audiovisual cues), and Agreeableness and Openness to Experience (for facial cues). Overall, our results concurred with findings from social psychology that had coded and inspected behavioral aspects related to audiovisual behavior and face expression in face-to-face interactions. Regarding the use of verbal content, **the numerous correlates found between all Big-Five traits and specific word category usage automatically computed from manual transcripts backed up findings previously documented in the literature in text blogs**.

Third, we addressed the task of predicting personality impressions using automatic cues. Our results concurred with previous attempts to automatically predict impressions in face-to-face interactions [Mairesse et al., 2007], radio broadcasts [Mohammadi et al., 2010], and multiparty meetings [Lepri et al., 2009], on **that nonverbal cues from audio and video are useful to predict the Extraversion trait only**. As in previous works, we found that prosodic cues were the audio cues with higher predictive power [Mairesse et al., 2007, Lepri et al., 2009], whereas among video cues, weighted energy images that aggregate visual activity throughout the video also provided competitive performance. **Facial expression cues and smile were also useful to predict the Extraversion trait**, with lower performance than audiovisual cues, which supports the idea that this trait can be predicted from this sources of visual information. In addition, we found that Agreeableness could be predicted better than with other nonverbal cues, which suggests that the vloggers' face may encode more information regarding this trait. However, we observed that compared to audiovisual models, the performance of facial expressions models showed larger variance across validation folds, which may result from having noisier

features. Finally, we found that **verbal content models were useful to predict Agreeableness, Conscientiousness, and Openness to Experience**. Specially for the case of Agreeableness, the result is important because despite being the second trait with higher reliability after Extraversion, the audiovisual models could not predict this trait and the performance of the facial expression model was poor. Similar results on the superior performance to predict these traits was found on analyzing verbal content from manual transcript on face-to-face interactions [Mairesse et al., 2007].

Despite the performance achieved for facial expressions, we found that some correlations between cues and impressions could not be easily explained, and that THR and HHM-based cues proposed provided slightly low performance with respect to a simple approach using basic statistics. As one possibly related issue, we do not know to what extent our results may be biased by vlogs showing very low activation (low PT) of facial expressions, which are numerous for the case of HMM. In addition, it is unclear how reliably facial expressions are estimated during speech, as some expressions such as surprise, sadness, or disgust are likely to be activated by the facial movement produced when talking. The superior performance of basic statistics may also motivate research work on alternative representations that exploit the distribution of features such as the use of Gaussian mixture models. Overall, we believe that our work in this chapter opens the door to future investigations on applying affective analysis to social video.

In comparison with manual transcriptions, we found that **the performance of verbal content models decreases significantly when using automatic transcriptions due to errors introduced by the ASR system**. Future work may exploit the estimation confidences output by the ASR to filter out unreliably recognized verbal content, or use automatic keyword spotting instead of full ASR. It would also be interesting to evaluate the how the ASR based LIWC-Big Five correlates compared between the manual and the automatic transcription. However, it may just be that we need for ASR technologies to improve before we can start using them for this task.

Finally, our work showed how computational models can improve when combining different sources of information. While fusion did not help much with features inside the same modality, **the combination of different modalities, namely audiovisual and facial cues for Extraversion, and facial and verbal cues for Agreeableness were very useful to improve performance**. To the best of our knowledge, this is the first time that multimodal integration is shown to be beneficial in prediction tasks in social video.

To conclude, we acknowledge that our experiments would benefit of having more data. In particular, this could help in experiments with noisy features, such as facial expressions; in experiments with n-grams, where the dimensionality of the feature vectors was considerably larger than the number of documents; and in fusion, to attempt other feature fusion methodologies.

5 Mining Crowdsourced Impressions of Vloggers

5.1 Introduction

We address the problem of annotating conversational social video with respect to vloggers' personal and social traits. This type of annotations have been recently used to train supervised machine learning algorithms that can characterize people automatically. The work on predicting personality traits from verbal and nonverbal information in Chapter 4 is an example of this. In this context, human annotations are used as ground truth, because assessing human traits is a human perceptual task in nature, and because the use of "thin slices" of behavioral data has been documented as suitable for the study of personal and social constructs based on first impressions (see discussion in Chapter 2). In the case of social media, the use of human annotations can also be exploited to collect new data about specific multimedia entities (e.g. a vlog) that can augment or complement the information already available through metadata.

In this chapter, we explore the use of crowdsourcing to collect multifaceted impressions of our dataset of vloggers. Though the use of crowdsourcing has already been exploited for other human annotation tasks (see related work in Chapter 2), to the best of our knowledge, this constitutes the first attempt to crowdsource this type of personal and affective impressions from online video. The advantages of using crowdsourcing in this context are twofold. First, crowdsourcing is potentially a fast and affordable method to scale human annotation to the large amount of social video available online, under the assumption that the annotation outcome is reliable. Second, by using crowdsourcing we have access to a large and diverse pool of annotators that we would not have otherwise in a traditional annotation task. In our experiments, we show evidence that the demographic variety of annotators in crowdsourcing sites resembles the diverse online community that consumes content in YouTube.

The annotation of multifaceted impressions is also an opportunity to go beyond the individual focus of most social media literature that has investigated users' traits and states individually (see Chapter 2), and to examine the interplay between different facets of people documented in social psychology [Dion et al., 1990]. In this regard, we argue that the three facets annotated: personality, attractiveness, and mood, though not exhaustively, cover a broad range of

impressions that can be built on the basis of vloggers' nonverbal behavior, which is the main focus of this thesis. In addition, using a broad list of impressions we address a more realistic scenario in which people make a variety of impressions while watching videos, and enables us to explore prototypical impressions that may be more relevant to the conversational vlogging setting, that are data driven, and not limited to a small number of labels as it is done with very specific prediction tasks.

We summarize the main contributions of this work as follows:

- We design a new crowdsourcing experiment to collect interpersonal impressions from vloggers based on video watching. We use reliability analysis to demonstrate that the impression agreement achieved with crowdsourcing compares well to the agreement achieved using traditional methods in related literature.
- We present an analysis of multifaceted crowdsourced impressions of vloggers that revisits some results reported in social psychology literature (with experiments mainly done in lab settings) regarding four points: the agreement of judgments; the interplay between personality, attractiveness, and mood impressions; the influence of physical and nonphysical facets in overall impressions of attractiveness; and the role of gender in forming impressions.
- We propose a probabilistic framework to represent vloggers in the multidimensional space of impressions, using a bag-of-impressions representation and probabilistic topic models. We show that a standard Latent Dirichlet Allocation model applied on the bag-of-impressions identifies meaningful prototypical impressions that are data-driven and emerge from video watching.
- We revisit the problem of social attention and show that certain topics related to impressions socially interpreted as more positive, are associated to larger audience responses, as measured from various metrics available from YouTube metadata.
- Finally, we address the task of automatic prediction of "topic impressions" using automatic analysis. We introduce comments as a new data source that contains information about vloggers, and investigate as well the performances of models based on the nonverbal and verbal information sources introduced in previous chapters.

The rest of the chapter is organized as follows. In Section 5.2, we describe the process to crowdsource annotations. In Section 5.3, we explain the automatic processing of vlogs for annotation and the automatic feature extraction. Then, in Section 5.4 and Section 5.4.3, we report our analysis on work reliability and impression annotations respectively. In Section 5.5 we present the experiments on mining multiple impressions using topic models and the connections of topics with social attention. Then, we address the prediction of topical impression in Section 5.6 and conclude in Section 5.7. Parts of this chapter have been published in [Biel and Gatica-Perez, 2012a]. Sections 5.5 and 5.6 are unpublished.

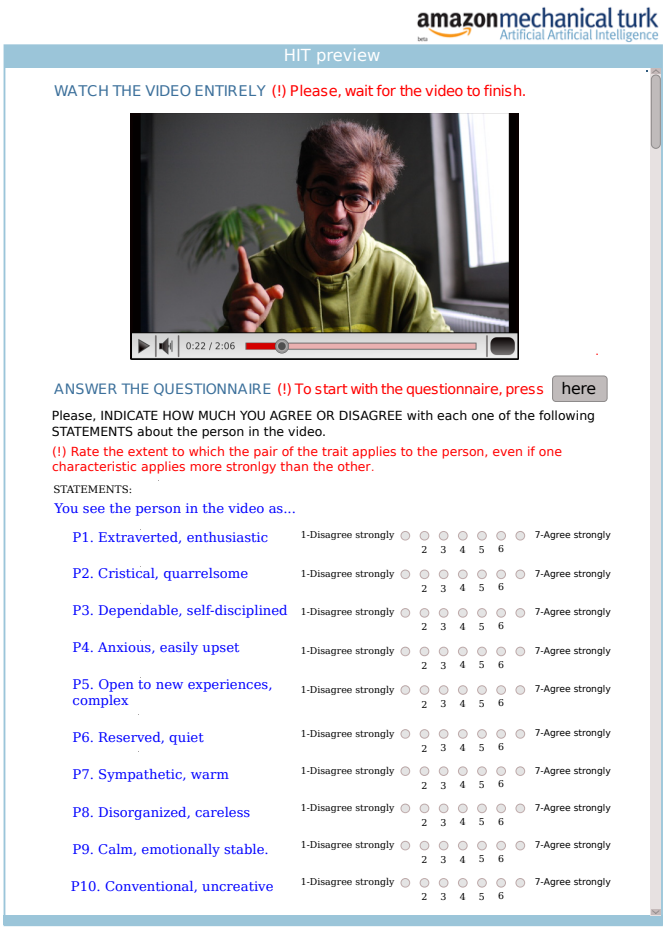

Figure 5.1: A view of the HIT designed to collect personality judgments from MTurk. On the top, the embedded vlog. On the bottom, the first of the four questionnaires used: the TIPI.

5.2 Crowdsourcing Vlogger Impressions

We crowdsourced the annotation of vlogger impressions using Amazon's Mechanical Turk. This task was inspired on the paradigm of video consumption in sites like YouTube, were people (the audience) are exposed to large amounts of content, and by playing bits of videos decide, on the basis of first impressions, what videos are worth watching and which ones to disregard.

Figure 5.1 shows a snapshot of the Human Intelligence Task (HIT) we designed, which consisted of two main components. The top part of the HIT contained an embedded video player to display the one-minute vlog slices obtained from preprocessing (see Section 5.3.1). The bottom part of the HIT included four questionnaires used to assess the personality, the attractiveness, the mood, and the demographics of vloggers. With the purpose of obtaining spontaneous impressions, we did not give any particular instructions to workers to fill the questionnaires apart form 1) watching the video entirely and 2) answering the questionnaires.

Questionnaire	Trait
Personality	Big-Five: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experience
Attractiveness	Beautiful, Likable, Friendly, Smart, Sexy, Overall attractiveness
Mood	Happy, Excited, Relaxed, Sad, Bored, Disappointed, Surprised, Nervous, Stressed, Angry, Overall mood
Demographics	Gender, Age, Ethnicity

Table 5.1: Summary of the crowdsourced annotations.

Table 5.1 summarizes the traits annotated using the four questionnaires, which we describe as follows.

Personality questionnaire

We annotated the personality of vloggers using the Ten-Item Personality Inventory (TIPI) designed by Gosling et al. [2003]. The TIPI measures the Big-Five traits of personality by means of 10 items (two items per scale on a 7-point likert scale), and is an instrument specially thought to be used when time is limited (it can be completed in approximately one minute). The questionnaire has shown reasonable psychometrics with respect to longer personality tests, and has already been used in several works to measure personality in social media settings [Gosling et al., 2007, Stecher and Counts, 2008].

In our case, we decided to use the TIPI in order to keep the annotation task as short as possible. The TIPI instructions were taken directly from Gosling et al. [2003], but were rephrased to ask workers about the vlogger personality. The Big-Five are summarized in Table 5.1. Our version of the form used can be found in the Appendix A.1.1.

Attractiveness questionnaire

We did not find any standard, short form in the literature to report attractiveness, and therefore we decided to design our own brief questionnaire. Our questionnaire was inspired on research investigating attractiveness from physical and non-physical facets [Fiore et al., 2008, Kniffin and Wilson, 2004]. First, we gathered a list of five facets that cover different aspects of attractiveness judgments, two facets of physical attractiveness (beautiful and sexy), and three facets of non-physical attractiveness (likable, friendly, and smart). Then, we phrased five items similarly to the personality questionnaire, using two adjectives to describe each facet, and a 7 point likert-scale. Finally, we added a sixth item to annotate the overall attractiveness of the vlogger. The full questionnaire can be found in the Appendix A.1.2.

Mood questionnaire

Standard mood forms such as the Profile of Mood States (POMS) [McNair et al., 1971] are designed to measure mood disorders rather than mood states in general conditions. Thus,

we decided to design a mood questionnaire based on existing social media research in mood. First, we took a long list of moods previously used in blogs research [Mishne, 2005], and simplified it to the twenty mood words that we identify as possibly being displayed by vloggers. Then, we use pairs of these words to define ten mood items that would cover a range of arousal and pleasure levels. The resulting questionnaire included three positive moods (from high to low arousal): excited, happy, relaxed; two neutral moods: bored, surprised; and five negative moods (from high to low arousal): anger, stressed, nervous, sad, and disappointed. Finally, we added an eleventh item to annotate the overall mood of the vlogger. This questionnaire is also presented in Appendix A.1.3.

Demographic annotations

We asked workers to annotate the gender, age, and ethnicity of vloggers. Though age and gender are optional metadata that can be extracted from YouTube user profiles, this information is often missing or unreliable. Apart from serving the purpose of collection, the annotation of these demographics represents an opportunity to measure the reliability and quality of the MTurk annotations, because these impressions are clearly more objective than the other questionnaires. We divided age and ethnicity in six categories each. For age, we considered: younger than 12, 12-17, 18-24, 25-34, 35-50, and older than 50. For ethnicity, we included: Caucasian, Black or African American, Asian/Pacific Islander, American Indian/Alaskan native, Hispanic, and Other, following standard US labels.

The actual design of the HIT resulted from an iterative process, in which we conscientiously refined it to discourage spammers from completing our tasks. With this purpose, we incorporated javascript and CSS to disable the HTML questions and control the flow of the HIT. First, the TIPI questionnaire was enabled only after the video had reached the end. Then, the attractiveness, mood, and demographic questions were enabled subsequently only after the previous questionnaire had been completed, to make sure that no question was skipped. Third, the final "Submit" button of the MTurk interface was hidden until all the questions were answered. Finally, in addition to the working time reported by MTurk, we logged the time MTurk workers spend on each of the components: time watching the video and time filling each questionnaire.

Gosling et al. [2003] suggested that the TIPI could be completed in one minute. Combining the video and the three other questionnaires, we estimated each HIT to take no more than three and a half minutes. In total, we posted 2,210 HITs to annotate five times each of the 442 vloggers. The HITs were restricted to workers with HIT acceptance rates of 95% or higher, and were limited to the US (1,768 HITs) and India (442 HITs), as these are the English speaking countries with more MTurk workers [Ross et al., 2010]. Before participating in our task, we asked MTurk workers to self-report their demographics (gender, age, and ethnicity), and to sign a consent of participation.

5.3 Vlog Preprocessing and Feature Extraction

5.3.1 YouTube Vlog Dataset and Preprocessing

With the purpose of bounding the time and costs that would take to annotate the full dataset, we limited the crowdsourcing task to a subset of vlogs by 1) randomly selecting one vlog per vlogger (to favor the variety of users), and 2) shortening the videos to a one-minute "slice". The use of "thin slices" is documented in psychology as suitable for the study of first impressions [Ambady and Rosenthal, 1992, Pentland, 2008]. In the case of personality, for example, research has suggested that few seconds are enough to make accurate impressions [Carney et al., 2007].

We used the automatic processing introduced in Chapter 3 to identify the conversational and non-conversational shots of vlogs. Then, we either shortened the first conversational shot or merged it with subsequent conversational shots when required, in order to obtain the first conversational minute of vlogs. In practice, we allowed durations between 50s and 70s so as to minimize the number of shot boundaries in the final vlog slice, and we discarded 27 vlog slices that were shorter than 50s after merging the conversational segments. The final dataset contained 442 vlogs of which 47% (208) corresponded to male and 53% (234) to female vloggers.

5.3.2 Automatic Feature Extraction

In Chapter 4, we explored verbal and nonverbal content of vlogs as potential sources of information to predict personality impressions. In this chapter, we introduce YouTube comments as a new source of information. Though we are not aware of any work providing a general picture of the type of comments found in YouTube, we argue that comments can potentially reveal information about vlogs and vloggers. For example, comments can refer to the vlogger discourse, to other previous comments, to the vlogger nonverbal behavior or looks, as well as to technical aspects from video editing or quality. In this context, a recent work on automatic categorization of video showed the use of comments to improve the performance of a baseline classifier using audiovisual analysis and metadata [Filippova and Hall, 2011], which supports the conjecture that comments contain useful information to characterize the video content (or the verbal content of vloggers, in our case).

Table 5.2 summarizes some basic statistics of the set of 442 videos and their comments. Note, that 70 videos did not have any comments at the time of data collection and therefore, these numbers are computed with respect to the other 372 videos. The median video had 14.50 comments and 256 words, compared to the 469 words of the median video transcription. However, the number of comments and words varies a lot across videos, as popular videos have very long comment threads. Considering only the most recent 200 comments for each video, the total text corpora from comments sums up to 328,207 words, which is 34% more data than the full transcription dataset (see chapter 4).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Comments	1	5	14.50	85.69	54.5	999
Words (all)	2	78	256	1609	997	37441
Words (200)	2	78	256	882.3	982	11035

Table 5.2: Summary statistics of comment threads in the subset of 372 vlogs with comments. Values did not change w.r.t. the full dataset. (All) indicates all available comments, (200) considers only the 200 most recent comments per thread.

As we did with transcriptions in Chapter 4, we explored two representations for comments using **LIWC** and **unigrams** (we did not use bigrams, because vectors become too sparse, which is a problem given the number of available samples). For each vlog, we created one single document with the text from its comment thread (limiting long threads to the 200 most recent comments). We processed comments to remove punctuation, to remove repeated letters in words (i.e. "aweeesooomeeee" was converted to "awesome"), and to stem words using Porter's algorithm. For LIWC, we used the 65 linguistic categories as features. For unigrams, we considered only n-grams that appeared in more than 10 documents, which resulted in 1286 unique words.

We observed the number of words found in the LIWC dictionary when processing comments increased from 67% to 84% after removing consecutive repeated letters. This percentage is still far from the 92% found when processing manual transcriptions with LIWC in Chapter 4, and indicates that despite the amount of data available in comment threads, the text is very noisy and it includes many typos/misspells that result from fast, spontaneous writing. In terms of unigrams, the numbers in comments compared poorly to the transcriptions, which basically indicate that are many words in our comment data that are unique.

5.4 Analysis of Crowdsourced Task

In this section, we discuss several aspects of MTurk task and the outcome annotations. In Section 5.4.1 we provide a basic discussion about the task completion and some basic description of the annotations. In Section 5.4.2 we analyze the reliability of the annotations, and in Section 5.4.3 we analyze the interplays between crowdsourced impressions.

5.4.1 Basic description

MTurk HIT Completion

The annotation tasks were completed by a total of 118 workers. The 1,760 HITs restricted to the US were completed by 91 workers and were finished within 12 days after being uploaded to MTurk, whereas the 442 HITs restricted to India were completed by only 27 workers and took 14 days to finish. This timing seemed reasonable given the 75h hours of expected work

Figure 5.2: Cumulative percentual distribution of MTurk annotations. Workers are ranked based on the number of HITs completed. The top ranked worker completed 17% of the HITs (400 HITs), 26 workers contributed with 80% of the annotations (1790 HITs), whereas 57 workers contributed with less than 5 HITs each (126 HITs).

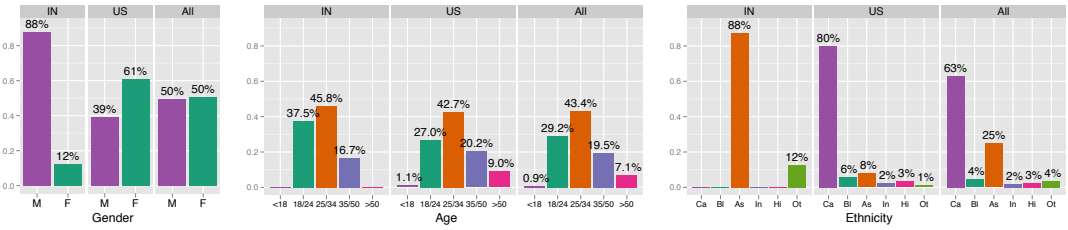

Figure 5.3: Demographic distribution (in %) of MTurk workers based on self-reported information (IN = Indian workers only, N = 24; US = US workers only, N = 89; ALL = overall sample).

(2min per HIT), the money spent on it, and the effort that would have been necessary to gather people offline to perform the task. However, it is slower than timings reported for other MTurk tasks [Mason and Suri, 2010]. Regarding individual work, each worker completed an average of 20 HITs. A two-tailed paired t-test on the distributions of HITs/worker showed no significant differences between the individual contribution of US and Indian workers ($t = -0.87$, $df = 103.021$, $p = 0.38$). However, as shown in Figure 5.2, the contribution varied substantially among workers, with one worker contributing to 17% of the annotations, and 26% of the workers providing up to 80% of the annotations. The average time of the TIPI questionnaire completion alone was 36.1s. Though this figure is substantially lower compared to the one-minute completion time suggested by Gosling et al. [Gosling et al., 2003], this result agrees with other recent studies in MTurk, where completion times of annotations were reduced with respect to experts' working time, which can be justified by the economic motive of MTurk workers [Soleymani and Larson, 2010].

MTurk Annotators Demographics

Figure 5.3 summarizes the demographics of MTurk workers. These numbers illustrate the ease of obtaining a significant demographically variate pool of annotators when using MTurk compared to gathering people offline. In our work, this diversity is desirable so as to better represent the variety of demographics found in online video audiences [1].

[1] http://www.youtube.com/advertise/demographics.html

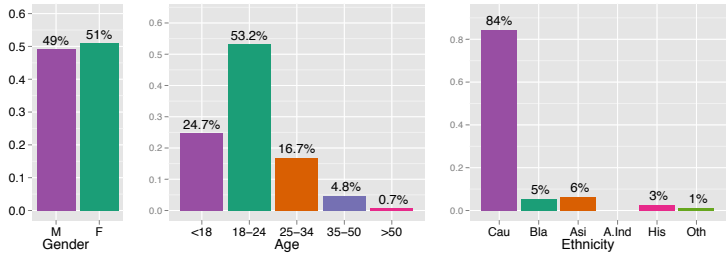

Figure 5.4: Vloggers demographics obtained based on the majority voting answers from crowdsourced annotations.

Our pool of annotators is balanced in gender but shows clear differences on the breakdowns between US (60% female and 40% male) and Indian workers (12% female and 87% male). Regarding age groups, we find most of MTurk workers between the ranges of 18-24 and 25-34. In addition, the typical Indian worker appears to be younger than the US workers. Finally, most of the US workers reported being Caucasian (80%), whereas, most of Indian workers reported themselves as Asian/Pacific Islander (88%). Interestingly, these demographic particularities resemble those reported by earlier investigations of the MTurk population demographics based on larger samples [Ross et al., 2010].

Crowdsourced Vlogger Demographics

Figure 5.4 shows the demographics of YouTube vloggers, which were obtained based on the majority voting answer of MTurk workers. Our sample of vloggers is mostly balanced in gender, and is mainly constituted by people between 18-24 and younger and from Caucasian ethnicity.

Though YouTube does not provide demographics of their video creators, these distributions compared to general demographics of YouTube users in what concerns to gender balance, and a prominent engaging among younger users. Unfortunately, we did not find any ethnicity demographic information regarding YouTube usage. In any case, our demographics are the mixed outcome of people that use YouTube, adopted videoblogging, and spend time making videos.

Impressions Statistics

As a first step towards understanding the type of impressions collected from MTurk, we computed a set of descriptive statistics (Table 5.3), including the mean, standard deviation, minimum, maximum, and skewness. As observed from the minimum and maximum scores, all the annotations span fully across the 7-points likert scale, which indicates that all the personality traits, attractiveness facets, and moods are found in the vlogging setting to some extent. The distribution of all personality traits and attractiveness facets are centered on the positive side of the likert scales (Mean ≥ 4) and showed little skewness (Skew $\leq \pm 1$), as it happens with positive moods (Happiness, Excitement, and Relax). In contrast, the rest of

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

Table 5.3: Basic descriptive statistics of vlogger impressions and Intraclass Correlation Coefficients ICC(1,k). All ICCs are significant with $p < 10^{-3}$.

moods (negative and neutral) are centered low on the negative part of the scale and result positively skewed (≥ 1). Independently of their ICCs, it is apparent that these moods are less frequent in vlogging, or that people in these states might be less likely to make a video.

The next step of our analysis is to evaluate the level of agreement that annotators achieve on their impressions from vloggers. The level of agreement can be interpreted as a measure of annotation quality that helps assessing whether crowdsourcing is a suitable setting for annotating online social video. To get an idea of the level of impression agreement among workers, we first measured the reliability on the three questions regarding the demographics of vloggers by means of the Fleiss' Kappa coefficient. Fleiss' Kappa assesses the reliability of categorical ratings and compensates for the agreement that could occur if raters were annotating at random. As one would expect from this type of annotations, the gender annotations showed high agreement ($\kappa = 91$). Clearly, the age and ethnicity annotations are more difficult to perform, yet they achieved a fair agreement for age ($\kappa = 29$), and moderate agreement for ethnicity ($\kappa = 46$).

5.4.2 Crowdsourced Judgements' Quality

We also evaluated the quality of the MTurk annotations in terms of the internal consistency of the personality test. The TIPI questionnaire is expected to have lower consistency compared to other personality inventories, as a consequence of using only two items per scale [Gosling et al., 2003]. However, very low consistency could indicate that MTurk workers are unreliably using the scales to fill up the personality questionnaires. To investigate this, we computed Cronbach's alpha reliability coefficient for each personality trait across all the annotations (N = 2,210), and report values in Table 5.3. We obtained alphas between .46 and .63 depending on the traits, a value range similar to that reported on the original TIPI report [Gosling et al., 2003], suggesting that MTurk workers are answering the questionnaires consistently from the point of view of the experimental design.

A core question of interest is to what extent workers are able to achieve any agreement on the basis of watching 1min slices of vlogs. In our setting, no agreement could result from two hypothetical situations in which either a) vloggers' behavior would not convey any impression information; or b) MTurk workers did not pay attention while completing the HITs. We computed the Intraclass Correlation Coefficients (ICCs) for each personality trait, as they are commonly used in psychology to measure the level of absolute agreement between annotators [Shrout and Fleiss, 1979]. Note that, in contrary to other existing reports of annotators agreement of personality [Gosling et al., 2007, Vazire and Gosling, 2004], we cannot use the ICC(2,k) measure, because each observer only annotated a subset of the data. Instead, we computed ICC(1,k) which is a measure of absolute agreement designed for experimental settings where each target is annotated by k judges randomly selected from a population of K judges, with $k < K$ [Shrout and Fleiss, 1979]. In our setting, we have $k = 5$ and $K = 113$. The last column of Table 5.3 shows the ICC(1,5) resulting of aggregating annotations across the 5 workers. Overall, the ICC(1,5) showed moderate reliabilities for all personality traits (.47 < ICC(1,5) < .77), attractiveness facets (.35 < ICC(1,5) < .69), and most moods (.48 < ICC(1,5) < .76) with the exception of Stressed (ICC(1,5) = .25).

Regarding the personality annotations, we now remark a few observations in the context of previous personality impressions research. The first one is that different personality traits achieved substantially different agreements. The second is that Extraversion is the trait achieving the highest level of agreement among observers. These two results have been repeatedly reported in research in personality [Gosling et al., 2007, Ambady et al., 1995], and are typically related to the evidence that the amount of observable information associated with some personality traits is larger than for others, and that this information varies with the context in which personality impressions are formed [Gosling et al., 2002, Borkenau and Liebler, 1992]. Third, most of the literature in face-to-face and video-taped impressions consistently reported Conscientiousness as the trait showing the second highest reliability. Thus, it is very interesting to see that in our case the trait achieving the second highest ICC is Agreeableness, and not Conscientiousness (which is in fact the trait with second lowest ICC). As argued by Gosling et al. [Gosling et al., 2002], who found a similar effect with the Openness

to Experience trait in personality impressions from bedrooms, this may well indicate that the vlogging setting is providing much more valuable information to form impressions of Agreeableness, as opposed to Conscientiousness.

We emphasize that the magnitude of the personality impression reliabilities compares well to other personality impression works, and indicate that overall there is substantial agreement on the personality impressions from MTurk. For example, Ambady et al. [Ambady et al., 1995] found that single personality impressions based on face-to-face interactions achieved a reliability between .07 and .27 for different traits, whereas Gosling et al. measured reliabilities between .23 and .51 for single impressions from bedrooms using the same TIPI questionnaire [Gosling et al., 2002]. Because these reliabilities were reported in terms of mean pair-wise correlations between raters, we computed these measures on our data for comparison (we considered only those annotators with more than 5 completed HITs). The resulting mean pair-wise correlations are: Extraversion (.44), Agreeableness (.36), Conscientiousness (.23) , Emotional Stability (.27), and Openness to Experience (.24). Other interesting works studying user profiles and websites reported reliabilities in terms of ICC(2,1) and may not be compared directly [Vazire and Gosling, 2004, Gosling et al., 2007].

Regarding the attractiveness and mood annotations, most of the judgments reliability was comparable to that of personality traits (see Table 5.3). Physical attractiveness facets such as Beautiful (.69) and Sexy (.61) achieved more agreement than non-physical facets like Friendly (.51) and Likable (.44). The overall attractiveness (.61) also achieved moderate agreement. Though we could not find any references in the literature to back up these findings, it is clear that non-physical impressions may require more information (e.g., longer observations) than first sight judgments of physical attractiveness. Interestingly, mood impressions were on average the impressions that achieved higher agreement compared to personality impression and attractiveness, with the exception of Nervous (.25). High arousal moods such as Happy (.76), Excited (.74), and Angry (.67), as well as the overall mood (.75) achieved highest agreement. This result is likely associated to the amount of visual and acoustic activity of these mood states compared to low arousal moods.

Overall, our experience with MTurk was that running a HIT required substantially more involvement than we had estimated in order to built a community of trusted workers. This included answering emails from workers, ensuring that workers understood the task and took enough time, and validating submitted HITs in order to built a community of trusted workers. In general terms, our experience using MTurk supports previous reports by others on the annotation of multimodal corpora [Soleymani and Larson, 2010]. Fortunately, this effort seems to be recompensed with annotations that have substantial agreement.

5.4.3 Analysis of Crowdsourced Impressions

In this section, we first study the interplay among the annotations of personalty, attractiveness, and mood. Then, we look at the impressions that elicit overall judgments of attractiveness and

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

Table 5.4: Pair-wise correlations of selected impressions. with ICC(1,k) > .50 (with the exception of absolute values lower than $r = .10$, all correlations are significant with $p < 10^{-3}$).

mood. Finally, we discuss some gender differences on impression formation.

Correlations between Impressions

We evaluated the extent to which vlogger impressions are associated to each other by means of pair-wise correlations (Table 5.4). For this analysis, we focus on traits that showed substantial agreement (we choose those ICC(1,k) > .50 arbitrarily), and we did not include overall attractiveness and overall mood, which we specifically address in the next subsection. We found a number of positive and negative effects that may be explained by a well-documented halo effect that suggests that attractive people are typically judged as holding more positive traits than unattractive people, with some exceptions [Dion et al., 1990]. For example, we found significant positive correlations between judgements of attractiveness and Extraversion (Beauty, $r = .20$, Friendliness, $r = .35$, and Sexiness $r = .17$), which have been previously reported in the literature in other settings [Borkenau and Liebler, 1992]. In addition, we found that Beauty is positively correlated with positive moods (Happiness, $r = .37$, Excitement $r = .33$, Relax $r = .25$), and negatively correlated with negative moods (Sadness, $r = -.15$, Boredom, $r = .-18$, Stress $r = -.14$, and Anger $r = -.15$). This halo effect may as well be mediating some of the correlations between Extraversion and moods (Happiness, $r = .47$ or Stress, $r = -.28$). It is important to highlight that, compared to Extraversion, Agreeableness shows even stronger correlations with attractiveness and mood (e.g. Beauty $r = .30$, Friendliness, $r = .57$, Hapiness $r = .38$, Anger $r = -.58$), which to the best of our knowledge may have not been observed in the literature because Agreeableness typically achieves less agreement in scenarios different than vlogging [Borkenau and Liebler, 1992]. Importantly, note that judgements of Extraversion and Agreeableness are not correlated ($r = .04, p = .30$). Finally, it is worth commenting the different associations between Relaxed and Extraversion ($r = -.12$), and between Relaxed and Agree-

ableness ($r = .40$), which is the only mood that shows opposite sign effects with these traits. Likely, in the first case, Relaxed was interpreted as calmed (opposite to excited), whereas in the second case it may have been judged as pleasant.

Overall Impressions of Attractiveness and Mood

We investigated the formation of overall attractiveness and mood impressions based on their several faceted impressions. This is relevant in order to identify the type of information that a single overall judgement would convey, if one attempts to simplify annotations to a few or a single item.

We used linear regression to test the contribution of physical and nonphysical attractiveness to the overall attractiveness impressions. Combining the physical facets of attractiveness alone explained 77% of the overall attractiveness variance ($R^2 = .77$, $\beta_{beauty} = .50, t = 27.7, p < 10^{-3}$,$\beta_{Sexy} = .22, t = 14.8, p < 10^{-3}$). Similarly, a model of nonphysical facets explained 44% of the overall attractiveness variance ($R^2 = .44$, $\beta_{Likable} = .38, t = 13.1, p < 10^{-3}$, $\beta_{Friendly} = .16, t = 5.8, p < 10^{-3}$, and $\beta_{Smart} = .22, t = 10.3, p < 10^{-3}$). Though in both cases we used stepwise linear regression procedures to evaluate different combinations of judgements, both final models included all the original facets. To test the contribution of the nonphysical facets on judging the overall attractiveness we compared the physical attractiveness model to a full model including all facets (physical and nonphysical) using an Analysis of Variance (ANOVA). The full model resulted to be significantly better than the physical attractiveness model ($F = 10.2, p < 10^{-3}$), indicating that nonphysical facets are also important on judging overall attractiveness, as it has been reported in the social psychology literature [Kniffin and Wilson, 2004]. The full model explained 80% of the attractiveness variance.

We repeated the linear regression experiment for the case of the overall moods. The linear model resulting of combining all moods explained 64% of the variance of the overall mood with main contributions from Happiness ($\beta_{Happy} = .33, t = 18.7, p < 10^{-3}$), Excitement ($\beta_{Excited} = .20, t = 12.7, p < 10^{-3}$), Relax ($\beta_{Relaxed} = .19, t = 15.3, p < 10^{-3}$) and Anger ($\beta_{Angry} = -.16, t = -10.8, p < 10^{-3}$), and small yet significant contributions from Surprise ($\beta_{Surprised} = .08, t = 6.9, p < 10^{-3}$) and Stressed ($\beta_{Stressed} = -.03, t = -1.9, p < 10^{-2}$). This is, moods with higher reliability seemed to dominate the association with overall mood.

Gender Differences on Impressions

We also explored whether there vlogger impressions differed depending on the gender of the annotators or the vloggers. For this purpose, we tested both rater's gender and vlogger's gender effects by means of one-way ANOVA for each characteristic independently, which we summarize in Table 5.5. Two-way ANOVA tests reported no interactions at all.

	Eff	Df	Sum Sq	Mean Sq	F value
Ext	R	1	29.1	29.0	15.0***
Agr	V	1	74.0	73.9	44.8***
Agr	R	1	267.5	267.4	162.2***
Cons	R	1	411.0	410.9	233.4***
Emot	R	1	528.9	528.9	286.5***
Open	R	1	83.6	83.5	55.5***
Beautiful	V	1	180.3	180.1	79.7 ***
Beautiful	R	1	32.4	32.3	14.3 ***
Likable	V	1	52.9	52.9	27.0***
Likable	R	1	254.3	254.2	130.1 ***
Friendly	V	1	48.2	48.21	26.0 ***
Friendly	R	1	317.9	317.9	172.0***
Smart	R	1	271.3	271.3	146.8***
Sexy	V	1	179.4	179.4	63.8 ***
Overall attr	V	1	118.6	118.6	55.1 ***
Overall attr	R	1	73.4	73.3	34.0 ***
Happy	V	1	35.5	35.5	13.0***
Happy	R	1	46.4	46.3	17.0 ***
Sad	R	1	51.8	51.8	228.1***
Bored	R	1	69.3	69.3	269.3***
Disapp.	R	1	30.2	30.2	108.5***
Surprised	R	1	4.91	4.9	18.1***
Nervous	R	1	498.1	498.1	194.5 ***
Stressed	R	1	49.5	49.5	216.9 ***
Angry	V	1	4.87	4.8	20.1 ***
Angry	R	1	26.21	26.2	108.3***
Over. mood	R	1	33.7	33.7	15.1 ***

Table 5.5: One-way ANOVA for gender effects. For space reasons only significant experiments are shown, Eff indicates the type of effect (R=Rater's gender effect, V= Vlogger's gender effect), Df = degrees of freedom, Mean Sq = Mean Square Error, Sum Sq = Sum of square, F= Fisher's F-ratio statistic.

Regarding personality, we found significant rater's gender effects on the annotation for all personality traits. In all cases, mean personality scores given by female raters were higher than scores given by male raters (mean values not reported for brevity reasons). In addition, we found vlogger's gender effects for Agreeableness only (female vloggers scored higher than male vloggers) and no interaction effects between vlogger and rater genders.

Regarding attractiveness annotations, we found significant effects of both rater's and vlogger's gender, and no interaction effects. In particular, we found rater's gender effect on judgements of attractiveness for all traits (in all cases female raters gave higher ratings of attractiveness than male raters), except for Sexual attractiveness. Vlogger's gender effects where significant for all facets of attractiveness (with female vloggers scoring higher than male vloggers) except for Smart. We also explored gender differences on the influence of nonphysical facets to the overall judgment of attractiveness, by replicating the linear regression of the previous section for annotators of the same gender. We found that for female annotators, adding nonphysical facets to physical facets had a stronger contribution to the overall attractiveness

rating ($F = 43.8, p < 10^{-3}$) than for males ($F = 29.03, p < 10^{-3}$).

Finally, we found rater's gender effects on all moods except Excitement and Relax. Male raters gave significantly higher scores than female for positive and negative moods. In addition, we found vlogger's gender effects for Happiness (female vloggers scored higher) and Anger (male vloggers scored higher).

To summarize, our analysis showed that several judgments are correlated to each other. However, it could also be that the links between traits are more complex than the linear associations investigated in this section. Next, we explore the use of a probabilistic approach to mine the crowdsourced annotations and discover topics that emerge from the interplays between them.

5.5 Mining Multifaceted Impressions With Topic Models

In this section we propose the use of the Latent Dirichlet Allocation (LDA) model to discover joint impressions of vlogger personality, attractiveness, and mood. With this, our goal is to model the interplay between impressions and to identify what basic aspects are useful to make overall distinct impressions from vloggers. We argue that these impressions may be useful aggregate traits to characterize vloggers.

LDA is a probabilistic generative model originally designed for discovering topical patterns in text documents based on word co-occurences [Blei et al., 2003]. The LDA model is also represented as a probabilistic graphical model in Figure 5.5. A document is generated by first sampling a distribution over topics θ_d. Then, for each word, a topic z is drawn and a word w_n is sampled from $p(w_n|z_n,\beta)$, i.e., the distribution of words conditioned to topic z. The distributions of documents over topics θ_d and the distribution of words β are learned during model inference [Blei et al., 2003]. The basic idea in LDA is that documents are represented as random mixtures over latent topics, and topics are characterized by a distribution over words. This is formally represented with the marginal probability of a document w as:

$$p(w|\theta,\beta) = \sum_z p(w|z,\beta)\,p(z|\theta) \tag{5.1}$$

where z denotes topic, θ is a distribution of documents over topics, and β is the parameter of Dirichlet prior in the per-topic word distribution.

LDA helps to explore document collections, by interpreting topics based on their most likely words and visually representing them using their most likely documents, or by characterizing documents with their most likely topics.

In order to use LDA for our data, we treated vloggers as documents, and impressions as words. In particular, we used two words (LOW and HIGH) to represent the each personality annotation and each attractiveness facets, and one word to represent each of the mood annotations. Then, we generated a document for each vlogger in which the frequency of each word was obtained

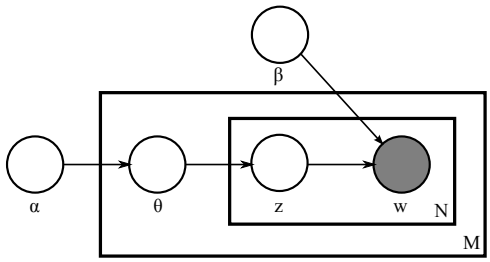

Figure 5.5: Graphical model representation of LDA. The boxes are plates representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

from the annotations given by the five raters. The frequency of the HIGH personality, HIGH attractiveness, and mood words were obtained as $f_w = \sum_{i=1}^{k} s_i(w) - k s_{min}$, were w is the word, k is the number of annotators per vlogger, $s_i(w)$ is the score given by annotator i when judging the trait associated with w, and s_{min} is the lowest possible score of the likert scale ($s_{min} = 1$). Similarly, the frequency of the LOW personality and LOW attractiveness words were obtained as $f_w = k s_{max} - \sum_{i=1}^{k} s_i(w)$, where s_{max} is the maximum possible score of the likert scale. Overall, we had 442 documents, a vocabulary of 33 unique words and 146,738 word tokens. Note that, as a result of this representation, all words appeared at least one time in every document.

We used Gibbs sampling to infer the distribution of over topics θ_d and the distribution of topics over words based on the our collection of 442 vloggers. The LDA hyperparameters were set to standard values ($\alpha = 0.1$, $\beta = \frac{50}{T}$) [Blei et al., 2003]. We explored the use of different number of topics, and we report results with $T = 6$.

5.5.1 Topic Interpretation

Figure 5.6 shows the top 7 words for each topic, together with their probability (the font-size of words is set proportional to the probability). It also shows the probability of each of the six topics $P(z)$. We noted that most topics resulted from a combination of personality, attractiveness, and mood impressions together, with the exception of Topic 4 and Topic 6 which are mainly characterized by attractiveness and mood impressions.

The figure also represents visually the topics on the basis of the 4 most probable vloggers for each topic. We clearly acknowledge that showing the snapshots together with the top words for each topic is a sensitive issue, not only for reasons of vloggers' privacy, but also because it could be interpreted as the researchers being judgmental of the vloggers themselves. This is not our intention at all. Rather, we found that the images were extremely powerful at illustrating the topical vlogger impressions,

Topic 1, Topic 2 and Topic 5 ranked multiple personality impressions among the most likely words. In particular, Topic 1 is characterized by low personality impressions of Conscientious-

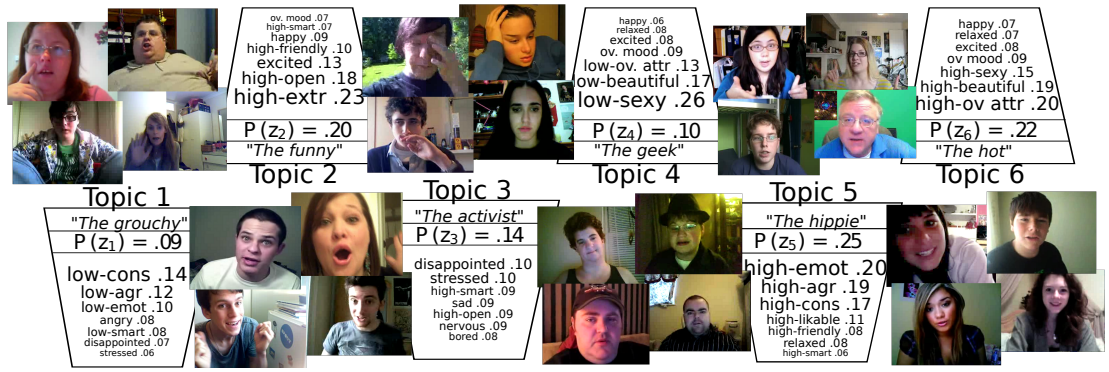

Figure 5.6: Discovered LDA topics. The titles on top of each topic are suggestions of "personae" that might capture the joint meaning of the top words. Topic descriptions and images are paired vertically.

ness ("low-cons"), Agreeableness ("low-agr"), and Emotional Stability ("low-emot). It also includes low judgements of intellectual attractiveness ("low-smart") and showing negative moods such as anger ("angry") or disappointment ("disappointed"). Topic 2 is characterized by high personality scores on Extraversion ("high-extr") and Openness to Experience ("high-open"). In addition, it includes high scores on nonphysical attractiveness ("high-friendly", "high-smart"), and impressions of positive moods ("excited", "happy"), and of overall mood ("ov. mood"). Compared to topic 1, Topic 5 is dominated by the high counterparts of the same personality traits: "high-emot", "high-agr", and "high-cons". It also included judgments of nonphysical attractiveness such "high-likable", "high-friendly" and positive mood "relaxed".

In contrast, Topic 3 represents a vlogger that is seen as disappointed, stressed, sad, and as a smart person ("high-smart"). This topic has "high-open" as the only personality judgment with high probability. This brings an interesting combination of a high score socially desirable on personality, and high scores on impressions of negative moods.

Topic 4 and Topic 6 are dominated by attractiveness and mood judgments. Topic 4 is characterized by low judgements of physical attractiveness ("low-sexy", "low-beautiful","low-over. attr"), but overall positive moods ("excited", "relaxed", "happy"). Finally, Topic 6 is characterized by high scores of physical attractiveness ("high-ov attr","high-beautiful","high-sexy"), and positive moods such as "ov. mood", "excited", and "relaxed".

As we expected, some of the correlations between impressions found in the previous section were captured by the LDA as word concurrencies in the topics. For example, we found that personality impressions from Conscientiousness, Agreeableness and Emotional Stability (which were correlated with $.41 < r < .64$) had words co-occurring in both Topic 1 and Topic 6, and did not co-occur with any other personality trait words in any other topic. In addition, Extraversion impressions, which correlated only with Openness to Experience ($r = .42$) co-occurred only with the latter in Topic 2. LDA also captured the main effect of physical facets on judgements of attractiveness impressions in Topics 4 and 6, with words of Sexiness and Beauty co-occurring with words of overall attractiveness. In the case of Topic 6, we found high physical

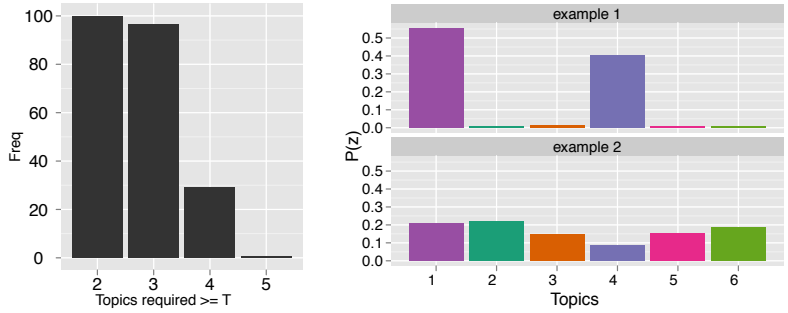

Figure 5.7: Left: percentage of documents that need more than T topics to cover 80% of their topic probability mass. Right: two examples of vloggers represented as a mixture of topics. The tops case is represented well by 2 topics, and the bottom is closer to a uniform distribution.

attractiveness together with positive moods, which may result from a positive halo effect. The model is also useful to capture some non-linear relationships that result from the interplay between personality, attractiveness and mood . For example, we find "excited" to co-occur with "low-beautiful" (in Topic 4) and "high-beautiful" (in Topic 6); or the words "happy" and "sad" to co-occur with "high-open". Finally, we found that specific word co-occurrences may add subtle connotations to specific judgements. Consider for example the overall impression of disappointment in combination with "angry" and "low-smart" (Topic 1) or displaying "sad", and "high-smart" (Topic 3).

Finally, we show that vloggers are indeed modeled as mixtures of topics, rather than being represented by one single topic. To measure this, for each vlogger, we counted the number of topics that accounted for 80% of the probability mass and plot the cumulative pattern in Figure 5.7. The plot shows that all vloggers need at least 2 topics to be characterized, and that 96% and 30% of them need at least 3 and 4 topics respectively. Figure 5.7 also shows two examples of a vlogger represented with 2 topics and another represented with 6 topics.

5.5.2 Vlogger Topics and Audience Response Metrics

One notable result of the characterization obtained using LDA is that some topics (e.g. 2, 5, and 6) cluster together impressions that can be socially interpreted as more positive than impressions clustered in other topics, because they correspond to high scores of personality impressions, judgements of high attractiveness, and positive moods. We hypothesize that these vloggers may be more appealing to interact with, and that consequently, this may result in significant differences in the amount of social participation that they generate with their vlogs in YouTube.

To test our hypothesis, we first selected a set of highly probable vloggers for each topic z, by thresholding their topic probability $P(z) > p_0$. Higher values of p_0 ensure that selected vloggers are highly ranked at the expense of obtaining a smaller sample of vloggers. In our experiments, we used $p_0 = .30$ to make sure that the sample sizes were large enough to measure

	Mean	SD	Skew
# views	288.91	8.18	0.98
# comments	12.50	5.31	0.99
# raters	14.71	4.71	1.27
av. rating	4.85	4.54	-0.76
# times faved	237 (53%)		

Table 5.6: Summary of YouTube metadata for the 442 vlogs. The # times faved variable is reported in number of vlogs.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
N	42	78	42	48	57	172
# views	146.64	633.45	105.62	72.34	251.21	373.16
# comments	0.60	8.97	0.82	0.53	2.84	7.26
# raters	10.95	22.52	5.91	6.04	11.33	12.92
av. rating	4.77	4.85	4.92	4.82	4.87	4.84
# times faved	0.65	0.80	0.35	0.25	0.50	0.60

Table 5.7: YouTube average metadata for each topic. The # times faved variable is reported in percentage of vlogs in each sample.

significant effects. Second, we test for any significant differences between the number of views, comments, times favorites, raters, and average ratings received by their vlogs. The number of views, comments, and raters follow power law distributions, and thus we use the logarithm to transform counts (base 10, after adding a start-value of 1). Because the median value of times favorite was 1, we transformed this variable to a binary value that measures whether the video was marked as favorite or not. Finally, the average rating, which ranges between 1 and 5, showed a large bias towards large values (mean = 4.85).This measure was transformed with a power of ten to reduce skewness. Table 5.6 summarizes all the response measures after transformations (mean and standard deviation values were transformed back to the original scale for interpretability).

Table 5.7 shows the numbers of views, comments, ratings, average ratings and proportion of vlogs favorites for each topic. The figures indicate that Topic 2, Topic 5, and Topic 6, receive higher number of views, comments, ratings, and are also favorited in larger proportions compared to other topics, supporting our hypothesis. We tested the differences between these measures by means of pair-wise T-tests, except for the proportion of vlogs favorites for which we used a Chi-squared test. To summarize, all the tests indicated significant differences between Topic 2, Topic 5, Topic 6 and the rest of the topics at $p < 0.005$, whereas only some tests were significant among these "positive" topics. For example, Topic 2 received significantly more views and raters than Topic 6 ($p < 0.005$) but not significantly more comments ($p = 0.7$). Regarding the average rating, t-tests showed that Topic 3 received a significantly higher average rating that the rest of the topics. Interestingly, this topic was a mix of positive personality scores and negative moods such as Disappointed or Sad. Overall, these results support the idea that vloggers that elicit more positive impressions also receive larger responses from the

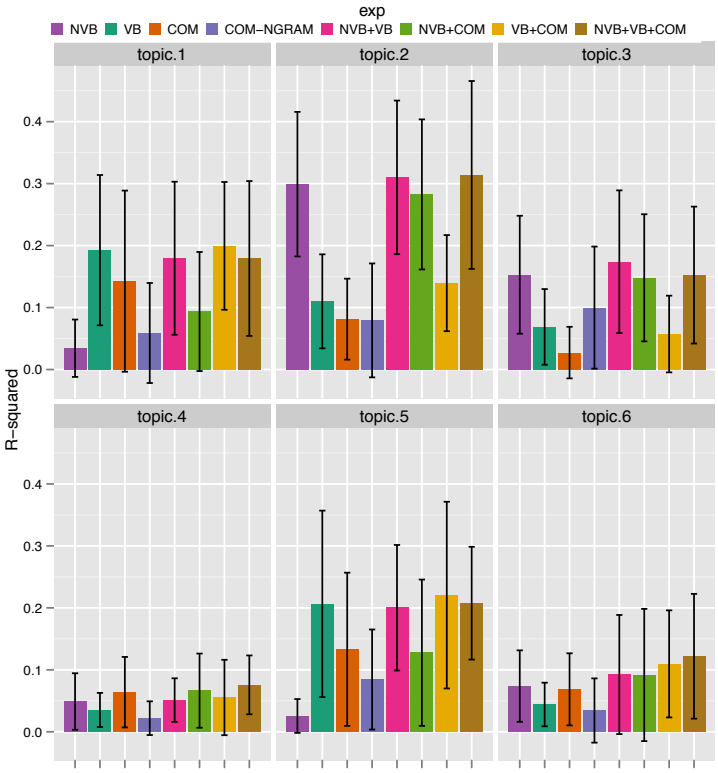

Figure 5.8: R-squared (R^2) prediction values of individual topic prediction tasks using audiovisual (AV), verbal content from transcripts (TRA), comments (COMs), and different combinations of them.

YouTube audiences.

5.6 Predicting Topic Impressions Automatically

In this section, we address the task of automatically predicting the topical impressions, discovered with LDA. Our experiments aimed to evaluate the extent to which vlogger topic probabilities can be predicted automatically compared to predicting traits individually, and to identify what sources of information carry useful information for this purpose. We approached this problem using one independent regression task per topic and taking topic probabilities as target scores.

We trained and tested different models using audiovisual features (AV), vlog transcriptions processed with LIWC (TRA), YouTube comments using both LIWC (COM) and unigrams (COM-ngram), and several combinations of them. With respect to the experiments in Chapter 4 the set of 408 vlogs was reduced to the 372 videos that have comments. In the experiments, we use a Random Forest model with features selected from a regularized Random Forest (RRF). To avoid overfitting on the selection, we use a 10-fold cross validation and run the RRF on the training data only for each fold (note that RF would not require a cross fold validation

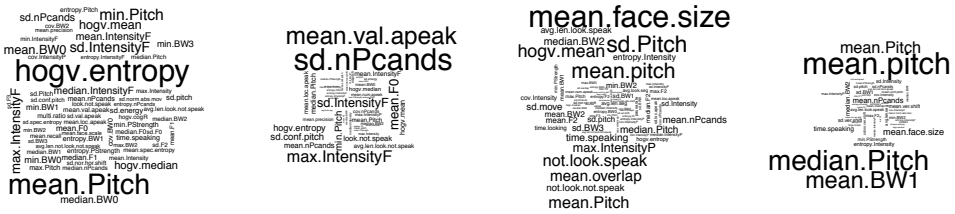

Figure 5.9: RF importance plotted as a tag cloud with AV model for Topic 2, Topic 3, Topic 4, and Topic 6 (from left to right).

otherwise). Results using inner selection with RRF slightly outperformed experiments using all features (i.e. no selection), so we decided to only report the former. Figure 5.8 shows R-squared (R^2) performance values for all topics and feature sets. We found that our experiments on personality prediction in the previous chapter, were valuable to interpret the results we present below.

The highest performance of a single feature set was achieved using AV for Topic 2 (R^2 = .30). With lower performance, AV was also useful to predict Topic 3 (R^2 = .10), Topic 4 (R^2 = .10), and Topic 6 (R^2 = .11). Looking at the topic representations of Figure 5.6, we found that the most likely words for Topic 2, "high-extr" and "high-open", are associated with the two best predicted personality traits using AV in Chapter 4. In addition, the third most likely word in Topic 2, "excited", is a high arousal positive emotion that can potentially be measured using the wMEI features (hogv.entropy, hogv.mean) and pitch (mean.pitch, min.pitch), as shown in Figure 5.9. The drop in performance in terms of R^2 of Topic 3, Topic 4, and Topic 6 cannot be explained on the basis of personality traits because Topic 3 has only one likely personality word. Instead, Topic 3 is described by high arousal negative emotion words such as "disappointed", and "stressed", which could be associated to some acoustic or visual activity. Figure 5.9 mainly relates this topic to voicing rate (mean.val.apeak) and pitch (sd.nPcands) related features. Topic 4 showed one main predictor associated to the distance to the camera (mean.face.size), and other cues such as not looking while speaking and not looking while not speaking. With respect to the distance, we hypothesize that a certain proximity to the camera may not be very appealing to audiences, but this needs further exploration in future work. For the multimodal patterns, it seems plausible that some of these behaviors may create poor impressions of vloggers. Finally, Topic 6 was mainly predicted from pitch features. This is explained with the trait being populated by females. Among the Top 10, 20, and 50 vloggers in Topic 6, we found 8, 19, and 43 females respectively.

In general, topics not predicted with the AV model, were found to be predictable using the TRA model. In particular, the highest prediction with TRA was achieved for Topic 5 (R^2 = .22) and Topic 1 (R^2 = .20), and to lesser extent for Topic 3 (R^2 = .07). Compared to AV, Topic 2 (R^2 = .13) and Topic 6 (R^2 = .08) were predicted by TRA with lower performance. The performance of TRA for both Topic 1 and Topic 5 can be explained by the presence of Conscientiousness, Agreeableness and Emotional Stability among the topics' most likely words, which are the traits

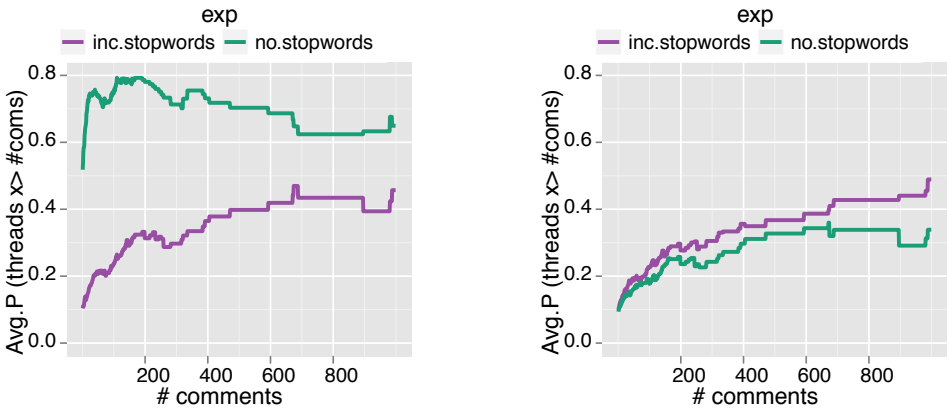

Figure 5.10: Average precision retrieving comment threads for transcripts: (left) Using unigrams; (right) using LIWC. Precision based on unigrams after removing stop words indicates similarity between transcripts and comment threads. For LIWC, precision falls due to the high similarity between all comment threads and transcripts once represented by LIWC categories.

best predicted by speech transcriptions in previous experiments (see Chapter 4). In particular, Topic 1 in Figure 5.6 captured the low scores of these traits ("low-cons", and "low-agr", and "low-emot"), and Topic 5 captured the high counterparts ("high-cons", and "high-agr", and "high-emot"). As with the AV model, Topic 3 and and Topic 6, were the ones showing lower results.

The results using YouTube comments showed lower performance than using AV and TRA. Using COM, best results were achieved for Topic 1 (R^2 = .14), and Topic 5 (R^2 = .11), and to lesser extent Topic 2 (R^2 = .08) and Topic 6 (R^2 = .08). Using COM-ngrams, best results were achieved with Topic 3 (R^2 = .10), Topic 2 (R^2 = .08) and Topic 5 (R^2 = .08). With the exception of COM-ngram and Topic 3, the results using verbal content from comments seem to show a similar trend to results using verbal content from transcriptions, specially if we compare the performance of Topic 1, Topic 2, and Topic 5 with respect the rest of the topics in Figure 5.8. This is relevant, because at this point, it remains uncertain what information from comments is actually used to learn impressions from vloggers. One possible explanation is that commenters directly provide impression information on their comments. Another explanation is that comments and transcripts contain similar information, which could be the case if people tend to comment on what vlogger said in the video.

We tested the second hypothesis using an information retrieval approach by measuring the similarity between transcripts and comment threads. To do so, we represented both types of documents with term vectors (unigrams), and computed the cosine similarity matrix resulting from all pair-wise combinations of transcripts and comment threads. Then, we compared the distances between corresponding transcripts and comment thread pairs with respect to the rest of the threads. For each transcript, we ranked all comment threads by decreasing similarity and measured the retrieval precision as $p = 1/k$, where k is the position were the

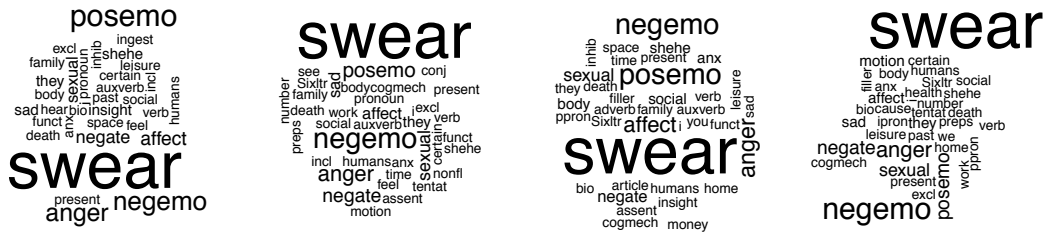

Figure 5.11: RF importance as a tag cloud using LIWC model for topics 1 and 5. From left to right: TRA model/topic 1, TRA model/topic 5, COM model/topic 1, COM model topic/5. The size of words was normalized with respect to the top predictor.

relevant document was found. Using this approach, very high precision indicates that a transcription and its corresponding comment thread are very similar compared to the rest of the threads. Figure 5.10 shows the average precision achieved with respect of the comment thread length (in number of comments). In addition to report the average precision $\hat{p} = \frac{\sum_i p_i}{N}$ (where N is the number of transcriptions), we also report the average pair-wise similarity between transcripts and corresponding comment threads ($\hat{s_p}$), and between transcripts and other comment threads ($\hat{s_o}$).

As shown in Figure 5.10 (left), the average precision when retrieving comment threads was low ($\hat{p}$ = .36) and the similarity was moderate ($\hat{s_p}$ = .43, $\hat{s_o}$ = .37). We also observed that longer comment threads were more similar to transcripts, which could be explained by stopwords dominating the unigram representation. Interestingly, we found that removing stop words increased dramatically the precision to $\hat{p}$ = .69 and that the difference between $\hat{s_p}$ and $\hat{s_o}$ increases substantially ($\hat{s_p}$ = .27, $\hat{s_o}$ = .08). This result supports the idea that documents and transcripts contain similar verbal content, and in part, explains why comments were found to be useful to classify the content of videos in a related work [Filippova and Hall, 2011].

However, in spite of the similarity between comments and transcripts, the COM and COM-ngram models did not provide results comparable to those of TRA. For the latter, we argued that our experiments using n-grams suffer from not having enough data. For the former, we also recomputed the cosine similarity between comments and transcripts based on the LIWC representation to understand how much of the similarity between transcripts and comments hold. After processed by LIWC, Figure 5.10 (right) shows that the average precision drops with ($\hat{p}$ = .34), and without stop words ($\hat{p}$ = .30). However, we note that both $\hat{s_p}$ and $\hat{s_o}$ become very high in both cases ($\hat{s_p}$ = .94, $\hat{s_o}$ = .93, $\hat{s_p}$ = .63,$\hat{s_o}$ = .68),which could be due to the implicit dimensionality reduction of LIWC.

We also compared the importance of predictors in TRA and COM models. Figure 5.11 shows the normalized importance of LIWC categories for TRA and COM and for Topic 1 and 5, which were the tasks that achieved higher performance. For both topics, TRA and COM shared the top predictors: swear, anger, posemo, negemo, and affect for Topic 1, and swear, anger, negemo, posemo, negate, and sexual for Topic 5.

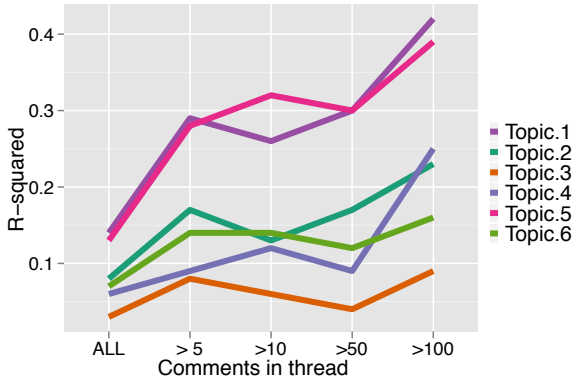

Figure 5.12: Variation of the R-squared performance with the number of comments in thread.

One of the limitations of our experiments is the amount of data available. This has a clear effect for the ngram model (COM-ngram), as the number of features is larger than the number of samples used for training. However, there is also an effect of some comment threads being very short, which may result in rather poor representations. To test the effect of comment thread length in the prediction of topical impressions, we run experiments on different subsets of data by by training and testing models using comment thread lengths of 5, 10, 50, and 100 comments different subsets. As shown in Figure 5.12, the performance of the predictor increases substantially, specially for Topic 1 (up to $R^2 = 42$) and Topic 5 ($R^2 = 39$), which are the topics best predicted with the TRA model. Clearly, longer aggregated comments provide more information than shorter comments, and this may have implications when training the model.

Finally, with the exception of the NVB-VB model, model combination did not result in significative improvements with respect to best single models.

5.7 Conclusions

We presented an original investigation on crowdsourced human impressions on a dataset of conversational vlogs, which includes five contributions. First, we showed that MTurk annotators agree substantially on their impressions of Big-Five personality traits, and many attractiveness facets and moods, which indicates that it is feasible to crowdsource this type of annotations in sites such as Mechanical Turk. In this context, however, it is still unclear whether the low reliability measured for some traits, attractiveness facets, and moods is due to their poor manifestation in conversational vlogging, because they are difficult to annotate on the basis of thin-slices, or both. Future work regarding the annotation task could study the reliability of annotations with respect to the slice duration. This is important, not only to assess whether the annotation reliability can be increased for some traits, but also to determine the minimal duration required for human observation, which is key if we aim to scale the annotation to thousands of vlogs. Along this line, it would also be interesting to evaluate

how many workers are required to reliably annotate each trait, and also explore annotation mechanisms that could determine the number of required annotations depending on the subject.

As a second contribution, our work adds to existing research of interpersonal perception in social media by investigating impressions formation in vlogging beyond personality traits, with the inclusion of attractiveness, and mood. Our analysis provided insights on the interplay among impressions and some gender differences, all supported by existing literature. Overall, we believe that the amount of behavioral data available in YouTube, could help to back up other findings from social psychology at a scale not done before. In the particular case of vlogging, future work may explore the annotation of other personal or social skills different than the ones explored here, and that may be very relevant to conversational social video: persuasion, influence, story-telling, entertainment, emotion, etc. These may also lead to the investigation of specific vlogging types, where vloggers develop different roles in marketing, product reviews, how-to vlogs, etc.

Third, we investigated the use of a bag-of-impressions representation and LDA to mine multi-faceted annotations, finding meaningful representations of vloggers. Regarding this characterization, it would be interesting to evaluate the relevance of the most likely topics for each vlogger on the basis of human annotations. The topical characterization of vloggers, could also be useful for vlogger retrieval, which may be a relevant task in certain scenarios or future applications. Future work could also explore the use of LDA or similar methods to mine an open (unrestricted) vocabulary of vlog first impressions.

Fourth, we showed that topics related to positive impressions are associated to vlogs with larger audience responses in YouTube. Though this result was first found for the case of personality traits in Chapter 4, the experiments here show that the actual relation between social attention and vlogger behavior is complex, and that is certainly better explained using multifaceted impressions. Overall, the results illustrates the importance of investigating the types of settings, behaviors, and content that mediate vlogging creation and consumption.

Finally, we addressed the prediction of topical multifaceted impressions for the first time, going beyond the tasks of personality or mood prediction recently studied in the social media literature. Our experiments showed that different sets of features: audiovisual analysis (AV), transcriptions (TRA), and YouTube comments (COM), were useful to predict different sets of topics, and that some results could be explained on the basis of our previous research predicting personality. We also found that the performance of YouTube comments could be explained by the similarity between comment content and transcripts and that both TRA and COM models were able to predict the same topics and shared some of the top predictors. We also observe that a drop in performance of COM with respect to TRA could be explained by having very short comments, and that taking comment threads longer than 5 comments substantially increased performance. Finally, we saw that combining feature sets achieved comparable performance to best single sets, with the advantage of not having to use prior

knowledge on the predictions used.

We believe that the prediction results using comments should be taken with care and be backed up with more data in future studies. This result could have many implications for the modeling of conversational social video, because comments could potentially replace the automatic analysis of vlogger verbal content in some settings, specially given the current performance showed by automatic speech recognizers. Having more data may also enable the development of other representations of verbal content, such as ngrams, that may exploit the spontaneous verbal content of comments and also its multilingual nature.

6 Conclusions

Mining conversational social video is a new domain for multimedia research. This video genre, available online in huge quantities, is a unique scenario for the study and characterization of complex human behavior in social media - of both vloggers and their audiences - that before our work had remained unexplored. In this thesis, we address three aspects of mining vlogs: the use of audio processing and computer vision techniques applied to conversational vlogs; the inference and understanding of some interpersonal and social processes that link vlogger behavior and vlog consumption in YouTube; and the computational modeling of vlogger behavior to automatically predict interpersonal impressions using machine learning techniques.

The rest of this chapter is organized as follows. In Section 6.1, we review the main contributions of each chapter. In Section 6.2 we discuss the limitations of our work. We conclude in Section 6.3 with future work.

6.1 Summary of contributions

In Chapter 3, we investigated the use of audio processing and computer vision techniques to analyze a dataset of conversational vlogs from YouTube. We proposed a principled method to segment vlogs and to extract audiovisual nonverbal cues from conversational shots. The audiovisual characterization of vloggers was used to investigate the links between vlogger behavior and the level of attention that videos receive in YouTube. Our study **showed evidence that some audio, video and multimodal cues extracted automatically are correlated with the average log-number of video view counts**, which indicates that, in addition to content, the nonverbal behavior of vloggers also plays a role in the communication and interpersonal perception processes involved in watching vlogs. To the best of our knowledge, this was the first time that this type of connection was investigated in social video. In addition, our analysis on a sample of videos showed that video edition elements are scarce in vlogging compared to

other types of online video, supporting the idea that this type of format is mainly driven by a communicative intent.

In Chapter 4, we approached the problem of interpersonal impressions with a focus on personality research and automatic behavioral analysis. First, we used personality impressions to revisit the problem of social attention and inspect to what extent vlogger personal traits can explain the links between behavioral cues and average view counts found in the previous chapter. Our analysis showed positive **linear associations for Extraversion, Openness to Experience, and Conscientiousness with respect to several social attention measures**, which indicates that people perceived as having high scores of these traits are more appealing to be watched. In contrast, **the Agreeableness trait showed a U-shape relation with attention**, indicating that vloggers on both extremes of this scale receive comparable treatment from audiences in terms of attention. These results concur with interpersonal perception theory on the type of personal traits that relate to people achieving attention or reacting to it in face-to-face interactions.

The automatic behavioral analysis was used to investigate what different sources of vlogger behavior can explain the personality impressions built from vloggers during video watching. Neither the acoustic nor the facial and body aspect of users had been addressed before in the context of personality impressions and social media research. The numerous personality correlates found in our analysis were backed up with findings from social psychology related to voice, face, and gesture in face-to-face interactions, and also by research analyzing verbal content in text blogs.

The correlation analysis results were consistent with other results on automatically predicting personality impressions. The performance achieved in our experiments concurred with previous attempts to automatically predict impressions in face-to-face interactions [Mairesse et al., 2007], radio broadcasts [Mohammadi et al., 2010], and multiparty meetings [Lepri et al., 2009], on that **nonverbal cues from audio and video are useful to predict the Extraversion trait mainly**, and that prosodic cues are the audio cues achieving the highest performance. Features that aggregate visual activity throughout the video, were the best performing features on the visual side. **Facial expression cues and smile were also useful to predict the Extraversion trait (with lower performance than the other audiovisual cues) and Agreeableness**. However, compared to audiovisual models, the performance of facial expressions models showed larger variance across cross-validation folds, which may result from having noisier features.

Verbal content models predicted Agreeableness, Conscientiousness, and Openness to Experience. Specially for the case of Agreeableness, the result is important because, despite being the second trait with highest reliability after Extraversion, the audiovisual models could not predict this trait's impressions and the performance of the facial expression model was poor. Similar results were found on analyzing verbal content from manual speech transcripts of face-to-face interactions [Mairesse et al., 2007]. However, we found that the performance of verbal content models decreases significantly when using automatic transcriptions due to

errors introduced by the ASR system. This indicates that the superior performance of verbal content to predict some personality traits may be penalized with the practical impossibility of building fully automatic models of these traits until the automatic transcription of online video improves. **Finally, we showed that computational models can be improved with the combination of different modalities, namely audiovisual and facial cues for Extraversion, and facial and verbal cues for Agreeableness**.

In Chapter 5, we proposed the use of crowdsourcing to collect human impressions from conversational social video. **Our analysis on the annotation reliability achieved for personality, attractiveness facets, and moods, suggests that it is feasible to crowdsource these annotations** in platforms such as Mechanical Turk, for the purpose of enriching multimedia datasets with personal and social constructs. The collected impressions allowed for the study of the interplay among several faceted impressions in social media, going beyond personality research. Specifically, we found that some of the correlations between personality, attractiveness, and mood, were explained by a positive halo affect documented in the social psychology literature. In addition, **we showed that the multifaceted characterization of vloggers via probability topic models is useful to enrich our understanding of the phenomenon of social media attention**, and also opens the door to build multifaceted perceptual computational models. Our experiments showed that audiovisual analysis and verbal content both from manual transcripts and comments are useful to automatically predict prototypical impressions of vloggers. However, results with comments need to be backed up with more data.

6.2 Limitations of the work

While our work contributed to several first findings about social video, it also has some limitations. The first limitation is the amount of data used, compared to the magnitude of data in other social media research today. Though it is clear that vlogs are available in YouTube and other platforms in large-scale, gathering vlogs in 2009 was not a simple task, mainly because the diversity the topics in vlogs results in a myriad of related keywords different than "vlog" or "vlogging" that make keyword search an ineffective way of retrieving vlogs, compared to other types of online video available. However, it seems reasonable to think that a number of successful approaches proposed on automatic video content categorization [Wang et al., 2010b, Filippova and Hall, 2011] could help to build larger collections of vlogs by automatically detecting conversational vlogs among general samples of online videos. The use of features related to the presence of speech and faces are candidates to help on this task. In addition, the use of crowdsourcing makes feasible the annotation of large samples of video if required to train and evaluate reliable vlog detectors. Nevertheless, it has to be mentioned that the amount of data used in our work, in terms of subjects and time duration, is larger than most current works in social computing done on audiovisual data by one order of magnitude.

The second limitation of our work is nonverbal cue validity, i.e., assessing that each nonver-

bal cue is appropriately capturing the aspect of conversational dynamics it is supposed to. Whereas most state-of-the-art audio processing and computer vision systems are usually benchmarked in standardized datasets, some of the techniques used in our work had not been tested for cue validity. This is true for our method to detect patterns of looking/not-looking, which was based on a frontal face detector instead of a face-tracker, or the CERT system, which despite having state-of-the-art performance, had not been tested before in noisy, amateur, conversational social video. Coding nonverbal cues manually is a laborious and expensive task and requires a substantial amount of time and expertise, and thus it is typically done in social psychology work with relatively small data samples [Knapp and Hall, 2005]. Despite the practical impossibility of validating nonverbal cues at large-scale, research in social computing has shown that imperfectly extracted nonverbal cues are effective to characterize human behavior [Pentland, 2008, Gatica-Perez, 2009]. In addition, these automatic feature extraction methods have often high cue reliability [Curhan and Pentland, 2007], i.e., multiple runs of the same feature extraction algorithm in one video provide the same exact feature values, which together with cue validity are desired requirements for behavioral measures in social psychology research.

The third source of limitation relates to the granularity and sparsity of metadata used to investigate the problem of social attention. The number of views and similar measures are useful indicators of the audiences reached by videos, specially when available in large numbers, but do not provide much information regarding the nature of the views (i.e., how, when, and why people watch a video), which limits the type of inferences that could be made in our work. For example, we do not know what percentage of views account for watching the video entirely, how many come from video discovering mechanisms internal or external to YouTube, or when in time the view took place with respect to the video upload time. Along this line, recent work from YouTube has shown the potential of analyzing fine-grained metadata to investigate several aspect of video consumption [Broxton et al., 2010, Brodersen et al., 2012]. While most of these metadata is confidential, other metadata such as the gender or the age of commenters, which are available online, can also be informative about the audience [Ulges et al., 2012] and could be useful to address research questions not contemplated here. In addition, the sparsity of data affects the significance of some findings such as the experiments using comments, because in our dataset the amount of comment data available varies substantially across videos. However, the problem could be overcome once we work with larger data collections.

6.3 Future work

There are many directions for extending the work done in this dissertation. Here, we discuss three main directions related to feature extraction, the use of crowdsourcing, and research on interpersonal perception.

While our approach was focused on the integration of technologies to mine vlogs, rather than

developing and improving the technologies themselves, future work could take vlogs as a test bed to improve specific feature extraction and video processing methods for behavioral cue extraction. For example, we found that background music and low audio quality were the two main causes of errors when processing the audio channel both for behavioral cue extraction and automatic speech recognition. While the quality of audio improves as people equip themselves with new technologies, it is difficult to imagine that some "poor" practices will disappear. Another problem related to feature extraction is the variety of positions and orientations of webcams that make looking/non-looking patterns difficult to estimate, as evidenced by our work. Given the importance of gaze in human interaction, it may be worth focusing in this very specific problem to improve the robustness of features computed from this modality. Future work could also investigate new ways to better characterize facial expressions of emotion. The superior performance of the basic statistic model compared to the two proposed methods based on segmentations suggests that richer representations of feature distributions such as Gaussian mixture models could capture the information in the face channel. In addition, research should investigate the effect of talking on the estimation of facial expression features, compared to estimates from silent facial expressions. Finally, work using large samples of data should revisit the n-gram representation to characterize verbal content and the predictive power of comments, both being open issues due to the amount of data available in our work.

Our work on using crowdsourcing to annotate multimedia corpora with social constructs could be directly extended to take into account the video duration required to annotate specific traits reliably. While the effect of "thin-slice" duration has already been investigated with respect to impression accuracy [Carney et al., 2007], the crowdsourcing aspect brings a new dimension to the problem because by minimizing video duration we can increase the number of samples annotated by workers. However, making sure that workers do watch the totality of any video shown is a practical open problem. Another mean to optimize crowdsourcing includes annotation modeling [Raykar et al., 2010, Whitehill et al., 2009] aimed to improve the aggregates of multiple judgments by modeling different aspects of annotation such as the annotator reliability or the task difficulty [Whitehill et al., 2009]. For example, it could be interesting to build a model that can identify what are the specific traits and vlogger individuals that are easy to annotate and which ones would benefit of having more annotations.

Finally, our approach to investigate interpersonal perception in vlogging can also motivate future work. First, research could investigate in detail some aspects of vlogger behavior that were not addressed in our work. For example, while we found evidence that the proximity to the camera creates different impressions of vloggers, we do not know what specific reactions from the audience are associated with vloggers being close to or far from the camera. Research could also go beyond the Big-Five traits, to investigate vlogger traits such as persuasion, influence, story-telling skills, entertainment skills, sentiment, satisfaction, etc. that may be relevant for specific vlogging applicationssuch as how-to, product reviews, or comedy videos, to mention some. In this regard, future work could study these specific formats, or start addressing other social video settings such as video testimonials. To conclude, research

could also leverage social video collections to investigate age group and gender differences at scale, or to analyze data from other social video repositories such as Youku.com (the chinese YouTube) to investigate cultural differences on both video creation and human perception in online video.

A An appendix

A.1 MTurk HIT Questionnaires

In this Appendix, we present the three questionnaires used to annotate vlogger impressions. In Section A.1.1 we present TIPI questionnaire Gosling et al. [2003], modified to ask about the vlogger personality. In Section A.1.2 and Section A.1.3 we present our attractiveness and mood questionnaires, respectively, which were designed as explained in Section 5.2.

A.1.1 Personality Questionnaire

Directions: Please INDICATE HOW MUCH YOU AGREE or DISAGREE with each one of the following STATEMENTS about the person in the video:

1- Disagree strongly, 2-Disagree moderately, 3-Disagree a little, 4-Neither agree nor disagree, 5-Agree a little, 6- Agree moderately, 7- Very much

Rate the extent to which the pair of traits applies the person, even if one characteristic applies more strongly than the other.

STATEMENTS:

You see the person in the video as...

P1. Extraverted, enthusiastic.

P2. Critical, quarrelsome.

P3. Dependable, self-disciplined.

P4. Anxious, easily upset.

P5. Open to new experiences, complex.

P6. Reserved, quiet.

P7. Sympathetic, warm.

P8. Disorganized, careless.

P9. Calm, emotionally stable.

P10. Conventional, uncreative.

A.1.2 Attractiveness Questionnaire

Directions: Based on your own judgement, please RATE HOW WELL each pair of ADJECTIVES describes the person in the video:

1-Not at all, 2-No, 3-Not much, 4-Neutral, 5-Somewhat, 6-Yes, 7-Very much

Rate the extent to which the pair of traits applies the person, even if one characteristic applies more strongly than the other.

ADJECTIVES:

A1. Pretty, handsome.

A2. Likable, nice.

A3. Social, friendly.

A4. Intelligent, smart.

A5. Sexy, hot.[1]

Finally, rate the overall attractiveness AND give your confidence about this specific rating:

1-Not attractive, 2-Moderately attractive, 3-Slightly attractive, 4-Neutral, 5-Slightly attractive, 6- Moderately attractive, 7-Very attractive.

A6. Overall attractiveness.

A7. Confidence of overall attractiveness rate (1-Not confident at all, 4-Neutral, 7-Very confident).

[1] Annotators were allowed to skip this if the vlogger was a minor or did not want to answer.

A.1.3 Mood Questionnaire

Directions: Based on your own judgement, please RATE HOW WELL EACH ADJECTIVE describes the MOOD of the person in the video:

1-Not at all, 2-No, 3-Not much, 4-Neutral, 5-Somewhat, 6-Yes, 7-Very much

Rate the extent to which the pair of traits applies the person, even if one characteristic applies more strongly than the other.

ADJECTIVES:

M1. Happy, glad.

M2. Enthusiastic, excited.

M3. Peaceful, relaxed.

M4. Sad, depressed.

M5. Bored, apathetic.

M6. Disappointed, dismayed.

M7. Surprised, amazed.

M8. Nervous, anxious.

M9. Stressed, worried.

M10. Angry, annoyed.

Finally, rate the overall mood AND give a confidence about this specific rating. Choose among the following:

1-Ver negative, 2-Moderately negative, 3-Slightly negative, 4-Neutral, 5-Slightly positive, 6-Moderately positive, 7-Very positive.

M11. Overall mood.

M12. Confidence of overall mood (1-Not confident at all, 4-Neutral, 7-Very confident).

Bibliography

E. Agulla, E. Rua, J. Castro, D. Jimenez, and L. Rifon. Multimodal biometrics-based student attendance measurement in learning management systems. In *Proc. IEEE International Symposium on Multimedia (ISM)*, 2009.

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

N. Ambady, M. Hallahan, and R. Rosenthal. On judging and being judged accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, 69(3):518–528, 1995.

M. Ashton, K. Lee, and S. Paunonen. What is the central feature of extraversion?: Social attention versus reward sensitivity. *Journal of Personality and Social Psychology*, 83(1):245, 2002.

K. Balog and M. de Rijke. Decomposing bloggers' moods. In *Proc. of the Int. Conf. on World Wide Web (WWW)*, 2006.

S. Basu. *Conversational scene analysis*. PhD thesis, Massachusetts Institute of Technology, 2002. Supervisor: Pentland, A.S.

L. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, tell me about yourself: Automatic assessment using short self-presentations. In *Proc. of Int. Conf. of Multimodal Interfaces (ICMI-MLMI)*, 2011.

F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and K. Ross. Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 5(4):30, 2009.

J.-I. Biel. Please, subscribe to me! analysing the structure and dynamics of the youtube network. Technical report, EPFL, 2009.

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

J.-I. Biel and D. Gatica-Perez. The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers. In *Proc. of ICWSM*, 2012a.

J.-I. Biel and D. Gatica-Perez. The YouTube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia*, 2012b.

J.-I. Biel and G. Gatica-Perez. Vlogsense: Conversational behavior and social attention in YouTube. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 7(1):33:1–33:21, 2011.

J.-I. Biel, O. Aran, and D. Gatica-Perez. You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2011.

J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. FaceTube: predicting personality from facial expressions of emotion in online conversational video. In *Proc. of. Int. Conf. of Multimodal Interaction (ICMI)*, 2012.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3: 993–1022, 2003.

A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis*, 23:257–267, 2001.

J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 2010.

P. Borkenau and A. Liebler. Trait inferences: Sources of validity at zero acquaintance. *Journal of Personality and Social Psychology*, 4(62):645–657, 1992.

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

A. Brew, D. Greene, and P. Cunningham. Using crowdsourcing and active learning to track sentiment in online media. In *Proc. of European Conf. on Artificial Intelligence (ECAI)*, 2010.

A. Brodersen, S. Scellato, and M. Wattenhofer. Youtube around the world: geographic popularity of videos. In *Proceedings of the 21st international conference on World Wide Web*, pages 241–250. ACM, 2012.

T. Broxton, Y. Interian, J. Vaver, and M. Wattenhofer. Catching a viral video. In *Proc. of IEEE Int. Conf. on Data Mining Workshops (ICDMW)*, 2010.

M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk. *Perspectives on Psychological Science*, 6(1):3, 2011.

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

E. A. Butler, B. Egloff, F. H. Wlhelm, N. C. Smith, E. A. Erickson, and J. J. Gross. The social consequences of expressive suppression. *Emotion*, 3(1):48–67, 2003.

D. R. Carney, C. R. Colvin, and J. A. Hall. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072, 2007.

M. Cha, H. Kwak, P. Rodriguez, and Y. Y. Ahn. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proc. of ACM Int. Measurement Conf. (IMC)*, 2007.

X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *Proc. of IEEE Int. Worksop on Quality of Sevice (IWQoS)*, 2008.

M. Cherubini, R. de Oliveira, and N. Oliver. Understanding near-duplicate videos: a user-centric approach. In *Proc. of ACM Int. Conf. on Multimedia (MM)*. ACM, 2009.

S. Counts and K. Stecher. Self-presentation of personality during online profile creation. In *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

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

N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2010.

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

G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. Sejnowski. Classifying facial actions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 21(10):974 –989, 1999.

J. Dovidio and S. Ellyson. Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. *Social Psychology Quarterly*, 45(2):106–113, 1982a.

J. F. Dovidio and S. L. Ellyson. 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, 1982b.

D. C. Evans, S. D. Gosling, and A. Carroll. 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)*, 2008.

B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36 (1):259 – 275, 2003.

K. Filippova and K. B. Hall. Improved video categorization from text metadata and user comments. In *Proc. of ACM SIGIR Int. Conf. on Research and Development in Information Retrieval (IR)*, pages 835–842, 2011.

A. Fiore, L. Taylor, G. Mendelsohn, and M. Hearst. Assessing attractiveness in online dating profiles. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2008.

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

A. Gill, S. Nowson, and J. Oberlander. What are they blogging about? Personality, topic and motivation in blogs. *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009a.

A. J. Gill, S. Nowson, and J. Oberlander. What are they blogging about? Personality, topic and motivation in blogs. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009b.

P. Gill, M. Arlitt, Z. Li, and A. Mahanti. YouTube traffic characterization: a view from the edge. In *In Proc. of ACM SIGCOMM IMC'07*, San Diego, CA, USA, 2007.

M. Goldhaber. Attention economics and the net. First Monday 2, 1998.

S. Gosling, S. Ko, and T. Mannarelli. A room with a cue: Personality judgments based on offices and bedrooms. *Journal of Research in Personality*, 82:379–98, 2002.

S. D. Gosling, P. J. Rentfrow, and W. B. Swann. A very brief measure of the big five personality domains. *Journal of Research in Personality*, 37:504–528, 2003.

S. D. Gosling, S. Gaddis, and S. Vazire. Personality impressions based on Facebook profiles. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2007.

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

T. Hain et al. Transcribing meetings with the amida systems. *IEEE Transactions on Audio, Speech and Language Processing*, 20(2):486–498, 2012.

J. A. Hall, S. D. Gunnery, and S. A. Andrzejewski. Nonverbal emotion displays, communication modality, and the judgment of personality. *Journal of Research on Personality*, 45(1):77 – 83, 2011.

M. Halvey and M. Keane. Exploring social dynamics in online media sharing. In *Proc. of the Int. Conf. on World Wide Web (WWW)*, 2007.

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

Hinton et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 2012.

G. Hitsch, A. Hortacsu, and D. Ariely. What makes you click: An empirical analysis of online dating. In *2005 Meeting Papers*, volume 207, 2005.

R. Hong, J. Tang, H.-K. Tan, S. Yan, C. Ngo, and T.-S. Chua. Event driven summarization for web videos. In *Proc. of ACM SIGMM workshop on Social media*, pages 43–48, 2009.

B. A. Huberman, D. M. Romero, and F. Wu. Crowdsourcing, attention and productivity. *Jour. Inf. Sci.*, 35(6), 2009.

H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proc. of Int. Conf. in Multimodal Interaction (ICMI)*, 2008.

F. Iacobelli, A. Gill, S. Nowson, and J. Oberlander. Large scale personality classification of bloggers. In *Proc of Affective Computing and Intelligent Interaction (ACII)*, 2011.

Y. Iizuka. Extraversion, introversion and visual interaction. *Perceptual and Motor Skills*, 1(74): 43–50, 1992.

D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez. Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proc. of. Int. Conf. of Multimodal Interaction (ICMI)*, 2008.

D. B. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech and Language Processing*, 17(3):501–513, 2009.

O. P. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2:102–138, 1999.

D. Kenny, C. Horner, D. Kashy, and L. Chu. Journal of personality and social psychology. *Consensus at zero acquaintance: replication, behavioral cues, and stability.*, 62(1):88–97, 1992.

F. Keshtkar and D. Inkpen. Using sentiment orientation features for mood classification in blogs. In *Proc. of IEEE Int. Conf. on Natural Language Processing and Knowledge Engineering (NLP-KE)*, 2009.

M. Kim, S. Kumar, V. Pavlovic, and H. Rowley. Face tracking and recognition with visual constraints in real-world videos. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

A. Kittur, E. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2008.

M. L. Knapp and J. Hall. *Nonverbal communication in human interaction.* Holt, Rinehart and Winston, New York, 2005.

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

B. Knutson. Facial expressions of emotion influence interpersonal trait inferences. *J. of Nonverbal Behavior*, 20(3):165 – 182, 1996.

A. Kramer. An unobtrusive behavioral model of gross national happiness. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2010.

A. D. I. Kramer and K. Rodden. Word usage and posting behaviors: modeling blogs with unobtrusive data collection methods. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2008.

G. Kruitbosch and F. Nack. Broadcast yourself on YouTube - really? In *Proc. of ACM Human-Centered computing (HCC)*, 2008.

B. Landry and M. Guzdial. Art or circus? characterizing user-created video on YouTube. Technical report, Georgia Institute of Technology, 2008.

P. Lange. Publicly private and privately public: social networking on YouTube. *Journal of Computer-Mediated Communication*, 1(13), 2007a.

P. G. Lange. Commenting on comments: Investigating responses to antagonism on YouTube. In *Society for Applied Anthropology Conf.*, 2007b.

B. Lepri, N. Mana, A. Cappelletti, F. Pianesi, and M. Zancanaro. Modeling the personality of participants during group interactions. In *Proc. of Int. Conf. on User Modeling, Adaptation, and Personalization*, 2009.

B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Employing social gaze and speaking activity for automatic determination of the extraversion trait. In *Proc. of Int. Conf. on Multimodal Interfaces (ICMI-MLMI)*, 2010.

G. Leshed and J. Kaye. Understanding how bloggers feel: recognizing affect in blog posts. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2006.

W. J. M. Levelt. *Speaking : from intention to articulation*. MIT Press, Cambridge, Mass., 1989.

J. Li and M. Chignell. Birds of a feather: How personality influences blog writing and reading. *International Journal of Human-Computer Studies*, 68(9):586–602, 2010.

W.-H. Lin and A. Hauptmann. Identifying ideological perspectives of web videos using folksonomies. In *AAAI Fall Symposium on Multimedia Information Extraction*, 2008.

G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (CERT). In *Proc. of IEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2011.

F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–501, 2007.

W. Mason and S. Suri. A guide to conducting behavioral research on amazon's mechanical turk. *Social Science Research Network Working Paper Series*, 2010.

R. R. McCrae and O. P. John. An introduction to the five-factor model and its applications. *Journal of Psychology*, 60:175–215, 1992.

D. McDuff, R. el Kaliouby, and R. Picard. Crowdsourced data collection of facial responses. *Proc. of Int. Conf. on Multimodal Interaction (ICMI)*, 2011.

D. McNair, M. Lorr, and L. Droppleman. *Profile of mood states (POMS)*. Educational and Industrial Testing Services, 1971.

M. Mehl, S. Gosling, and J. Pennebaker. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Jour. of Per. and Social Psych.*, 90(5):862, 2006.

R. Mihalcea and H. Liu. A corpus-based approach to finding happiness. In *Proc. of Spring Symposia on Computational Approaches to Analyzing Weblogs (CAAW)*, page 19, 2006.

G. Mishne. Experiments with mood classification in blog posts. In *Proc. of ACM SIGIR 2005 Stylistic Analysis Of Text For Information Access (SATIA)*, 2005.

A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of ACM Int. Measurement Conf. (IMC)*, San Diego, CA, USA, 2007.

A. Misra. Speech/nonspeech segmentation in web videos. In *International Speech Communication Association Proc. of IEEE Conf. of the Int. Speech Conference Association (Interspeech)*, 2012.

G. Mohammadi, A. Vinciarelli, and M. Mortillaro. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proc. of ACM Multimedia Workshop on Social Signal Processing (SSP)*, 2010.

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. *American Communications Journal*, 10 (2), 2008.

D. Nguyen and J. Canny. More than face-to-face: Empathy effects of video framing. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2009.

T. Nguyen, D. Phung, B. Adams, T. Tran, and S. Venkatesh. Classification and pattern discovery of mood in weblogs. *Advances in Knowledge Discovery and Data Mining*, 2010.

B. Ni, Y. Song, and M. Zhao. Youtubeevent: On large-scale video event classification. In *Proc. of IEEE Int. Conf. on Computer Vision (ICCV)*, 2011.

S. Nowson and J. Oberlander. Identifying more bloggers: Towards large scale personality classification of personal weblogs. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2007.

J. Oberlander and S. Nowson. Whose thumb is it anyway? Classifying author personality from weblog text. In *Proc. the Annual Meeting of the Assoc. for Computational Linguistics*, 2006.

S. O'Donnell, K. Gibson, Milliken, and S. J. M. Reacting to YouTube videos: exploring differences among user groups. In *Proc. of Int. Communication Association annual Conf.*, 2008.

R. D. Oliveira, M. Cherubini, and N. Oliver. Looking at near-duplicate videos from a human-centric perspective. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 6(3):15, 2010.

J. Pennebaker and L. King. Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1999.

A. S. Pentland. *Honest Signals: How They Shape Our World*, volume 1 of *MIT Press Books*. The MIT Press, 2008.

V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 99:1297–1322, 2010.

J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proc. of ACM SIG Int. Conf. on Human factors in Computing Systems (CHI)*, 2010.

D. Sanchez-Cortes, P. Motlicek, and D. Gatica-Perez. Assessing the impact of language style on emergent leadership perception from ubiquitous audio. In *Proc. on Int. Conf. of Mobile and Ubiquitous Multimedia (MUM)*, 2012.

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

P. Shrout and J. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proc. of SIGIR Workshop on Crowdsourcing for Search Evaluation*, 2010.

K. Stecher and S. Counts. Spontaneous inference of personality traits and effects on memory for online profiles. In *Proc. AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2008.

F. Steele Jr, D. C. Evans, and R. K. Green. Is your profile picture worth 1000 words? Photo characteristics associated with personality impression agreement. In *Proc. of AAAI Int. Conf. of Weblogs and Social Media (ICWSM)*, 2009.

A. Sureka, P. Kumaraguru, A. Goyal, and S. Chhabra. Mining youtube to discover extremist videos, users and hidden communities. *Information Retrieval Technology*, pages 13–24, 2010.

G. Szabo and B. A. Huberman. Predicting the popularity of online content. *Communications of the ACM*, 53(8):80–88, 2010.

S. Tong, B. Van Der Heide, L. Langwell, and J. Walther. Too much of a good thing? the relationship between number of friends and interpersonal impressions on facebook. *Journal of Computer-Mediated Communication*, 13(3), 2008.

A. Ulges, M. Koch, and D. Borth. Linking visual concept detection with viewer demographics. In *Proc. of ACM Int. Conf. on Multimedia Retrieval (IMR)*, 2012.

S. Vazire and S. D. Gosling. e-Perceptions: Personality impressions based on personal websites. *Journal of Research in Personality*, 87:123–132, 2004. ISSN 0022-3514.

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

J. B. Walther, B. Van Der Heide, S.-Y. Kim, D. Westerman, and S. T. Tong. The role of friends' appearance and behavior on evaluations of individuals on facebook: Are we known by the company we keep? *Human Communication Research*, 34(1):28–49, 2008.

S. S. Wang, S. Moon, K. H. Kwon, C. A. Evans, and M. A. Stefanone. Face off: Implications of visual cues on initiating friendship on facebook. *Computers in Human Behavior*, 26(2): 226–234, 2010a.

Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtubecat: Learning to categorize wild web videos. In *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010b.

M. Wattenhofer, R. Wattenhofer, and Z. Zhu. The youtube social network. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

M. Wesch. Youtube and you: Experiences of self-awareness in the context collapse of the recording webcam. *Explorations in Media Ecology*, 8(2):19–34, 2009.

J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, 22(2035-2043):7–13, 2009.

L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith. Visual memes in social media. In *Proc. of ACM Int. Conf. on Multimedia (MM)*, volume 11, page 53, 2011.

T. Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44:363–373, 2010.

M. Zancanaro, B. Lepri, and F. Pianesi. Automatic detection of group functional roles in face to face interactions. In *Proc. of Int. Conf. on Multimodal Interaction (ICMI)*, 2006.

X. Zhang, C. Xu, J. Cheng, H. Lu, and S. Ma. Effective annotation and search for video blogs with integration of context and content analysis. *IEEE Transactions on Multimedia*, 11(2), 2009.

M. Zink, K. Suh, Y. Gu, and J. Kurose. Watch global, cache local: YouTube network traffic at a campus network - Measurements and implications. In *Proc. of SPIE Multimedia Computing and Networking (MMCN)*, 2008.

Joan-Isaac BIEL

Born in Spain, Sep 1983
Single, holder of B Permit
Chemin du Reposoir 26, 1007 Lausanne
☎ (41)774671769
✉ joanisaac.biel@gmail.com

Profile I am a PhD candidate at EPFL. I am an independent and creative engineer with a passion for multimedia and social media applications using machine learning and behavioral understanding.

Education

PhD, Electrical Enginieering — Oct 2008 - May 2013 (expected)
Ecole Polytechnique Fédérale de Lausanne (EPFL) — Lausanne, Switzerland

M.S., Electrical Engineering — Sep 2001 - Jan 2007
Technical University of Catalonia (UPC) — Barcelona, Spain

Experience

Research Assistant — Oct 2008 - present
Idiap Research Institute - Social Computing Group — Martigny, Switzerland

During my PhD, I investigated the automatic modeling of human nonverbal behavior (NVB) in social media.

- I studied the link between NVB and social attention. I integrated a solution to automatically extract NVB cues from audio and video and tested the correlation between NVB cues and YouTube video views for more than 2,000 videos with significant effects. I reported my work in two scientific publications and was invited as speaker in two more workshops.
- I built a YouTube vlog collection that required manual validation. I developed a web app to play videos and aid manual labeling, and engaged colleagues to help me annotate 6,000 videos.
- Investigated the task of automatically predicting personality from YouTube vlogs.
- Wrote 7 scientific papers, 2 journals, and 1 book chapter on social media and multimedia. Participated in 5 conferences and was awarded a travel grant for one of them.
- Co-authorized a 6-months SNSF-funded (IM2) project and supervised a master student.

Research Intern — Jul 2012 - Aug 2012
Yahoo! Labs - Social Media Engagement Group — Barcelona, Spain

- I investigated on the retrieval of topical conversations around social media photo collections using automatic analysis of text and machine learning.

Visiting Resarcher — Apr 2012 - June 2012
HP Labs - Social Computing Group — Palo Alto, California

- I collected a set of 10,000 videos from YouTube and investigated the suitability of conversational analysis techniques on a variety of video types.

Visiting Researcher — Mar 2012 - Aug 2012
Intenational Computer Science Institute - Speech Group — Berkeley, California

- For my master thesis, I addressed the automatic identification of spoken languages. I collaborated with a postdoc on integrating a system to automatically classify audio files amongst 23 different languages. We participated in an international evaluation with 21 research teams and were nominated best amongst the 6 novice teams. I was also awarded High Honors for my thesis.
- I investigated on acoustic event detection for in-car applications sponsored by Volkswagen Electronics Lab. I wrote a conference paper, and wrote a project proposal extension that was successfully accepted by Volkswagen.

Research Assistant Feb 2006 - Feb 2007
TALP Research Center Barcelona, Spain

- I developed an acoustic event detection system for a smart meeting room. I implemented a solution to classify real-time audio segments from microphones amongst 16 possible sounds. I was invited to submit a report and a videodemo to a national spanish research contest targeted to master thesis, and was awarded finalist.

Marketing Assistant Sep 2005 - Jan 2006
ONO Telecomunicaciones - Customer Fidelity Department Barcelona, Spain

- I assisted a group of 4 people on analyzing the effects of marketing campaigns targeted to unsatisfied customers.
- Granted with a 9 months salary to visit the US as a visiting researcher by the AMIDA program

Training

Venture Challenge Feb 25th - May 27th 2010 (56h)
Entrepreneurship course organized by venturelab Geneva, Switzerland

Machine Learning Summer School Aug - Sep 2009 (2 weeks)
Cambridge University Cambridge, UK

Technical skills

Software development: Fluent programing in Python, comfortable with C/C++ and spontaneous use of Java.

Multimedia processing: Automatic audio, video, and text processing from multimedia and social media data.

Social computing: Design, development and evaluation of human behavioral experiments with social psychology foundations and computational techniques.

Machine Learning and statistical analysis: Good understanding and usage of machine learning algorithms and statistical analysis using R and Matlab.

Large-scale data analysis: Basic knowledge of MapReduce, HDFS, Hadoop streaming and Pig.

Honors and awards

Idiap Research Award December 2011
This award is given yearly to one Idiap PhD student for his outstanding publication record and high quality research.

MS Thesis High Honors January 2008
Technical University of Catalonia (UPC)

Scholarships by Technical University of Catalonia (UPC) 2001- 2006
Received 8 scholarships upon obtaining the highest grade in topics such as signal processing, mathematics, programming, and electronics.

Finalist and Honor award of the VI Certámen Arquímedes Dec 2007
Spanish Ministry of Education and Culture
This yearly contest is given to the best research works carried out for Master Students from all the universities in Spain.

AMIDA Trainee Feb 2007
Awarded with a European AMIDA training program grant for a 9 months internship as a visitor researcher at the International Computer Science Institute (Berkeley, CA).

Scientific activities

Invited scientific talks

Vlogcast Yourself: Exploring Nonverbal Behavior in Social Media. Idiap 20th Anniversary and IM2 Summer Institute, Martigny, Sep, 2011.

Vlogcast Yourself: Nonverbal Behavior and Attention in Social Media. Aizawa Yamasaki Lab, University of Tokyo, Tokyo, Nov. 2010.

Invited as reviewer

IEEE Transactions on Multimedia (2012), IEEE SMC 2012, AAAI ICWSM 2012, ACM Multimedia 2011, AAAI ICWSM 2011, ACM Multimedia 2010, IEEE ICME 2010, ACM MIR 2010, ACM WSM 2009,

Publications

Journals

J.-I. Biel and D. Gatica-Perez The YouTube Lens: Crowdsourced Personality Impressions and Audiovisual Analysis of Vlogs in *IEEE Transactions on Multimedia*, Vol. 15, No. 1, pp. 41-55, Jan. 2013.

J.-I. Biel and D. Gatica-Perez. "VlogSense: Conversational Behavior and Social Attention in YouTube" in *ACM Transactions on Multimedia Computing, Communications, and Applications*, Special Issue on Social Media, Oct 2011.

Book chapters

J-I. Biel and D. Gatica-Perez, "Call Me Guru: User Categories and Large-Scale Behavior in YouTube" in S. Boll, J. Luo, S. Hoi, and D. Xu (Eds.), *Social Media Computing, Springer*, 2011.

Conferences

J.-I. Biel, V. Vtsminaki, V., J. Dines and D. Gatica-Perez Hi YouTube! What verbal content reveals in social video, *submitted for a conference*, 2013.

J.-I. Biel, Teijeiro-Mosquera, L. and D. Gatica-Perez FaceTube: Predicting Personality from Facial Expressions of Emotion in Online Conversational Video in *Proceedings International Conference on Multimodal Interaction (ICMI)*, Santa Monica, Oct. 2012.

J.-I. Biel and D. Gatica-Perez The Good, the Bad, and the Angry: Analyzing Crowdsourced Impressions of Vloggers in *Proceedings of AAAI International Conference on Weblogs and Social Media (ICWSM)*, Dublin, June 2012

J-I. Biel, O. Aran, and D. Gatica-Perez, "You Are Known by How You Vlog: Personality Impressions and Nonverbal Behavior in YouTube" In *Proc. AAAI Int. Conf. . on Weblogs and Social Media (ICWSM)*, Barcelona, July 2011.

J.-I. Biel and D. Gatica-Perez, "Vlogcast Yourself: Nonverbal Behavior and Attention in Social Media" in *Proc. Int. Conf. on Multimodal Interfaces (ICMI-MLMI)*, Beijing, Nov. 2010.

J-I. Biel and D. Gatica-Perez, "Voices of Vlogging" In *Proc. AAAI Int. Conf. on Weblogs and Social Media (ICWSM)*, Washington DC, May 2010.

J-I. Biel and D. Gatica-Perez, "Wearing a YouTube Hat: Directors, Comedians, Gurus, and User Aggregated Behavior" In *Proc. ACM Int. Conf. on Multimedia (MM)*, Beijing, China, Oct. 2009.

Predoctoral

C. Müller and J-I. Biel, E. Kim, and D. Rosario, "Speech-overlapped Acoustic Event Detection for Automotive Applications" In *Interspeech*. Brisbane, Australia, 2008.

C. Müller and J-I. Biel, "The ICSI 2007 Language Recognition System" In *the Odyssey Workshop on Speaker and Language Recognition*. Stellenbosch, South Africa, 2008.

A. Temko, C. Nadeu, and J-I. Biel, "Acoustic Event Detection: SVM-based System and Evaluation Setup in CLEAR'07". CLEAR'07 Evaluation Campaign and Workshop, Baltimore, MD, USA, May 2007, in *Multimodal Technologies for Perception of Humans, LNCS,* vol. 4625, pp. 354–363, Springer, 2008.

Language skills

Catalan (native), Spanish (native), English (C1), French (B1)

Hobbies

I have been practicing yoga for over a year: it brings me mental and physical strength. I cook daily: it relaxes me and enables experimentation and creativity. I am a seasonal sportsman: I ski and snowshoe in winter, and run and hike during summer.