Hirability in the wild: Analysis of online conversational video resumes

Laurent Son Nguyen and Daniel Gatica-Perez

***Abstract*—Online social media is changing the personnel recruitment process. Until now, resumes were among the most widely used tools for the screening of job applicants. The advent of inexpensive sensors combined with the success of online video platforms has enabled the introduction of a new type of resume. Video resumes are short video messages where job applicants present themselves to potential employers. Online video resumes represent an opportunity to study the formation of first impressions in an employment context at a scale never attempted before, and to our knowledge they have not been studied from a behavioral standpoint. We collected a dataset of 939 conversational English-speaking video resumes from YouTube. Annotations of demographics, skills, and first impressions were collected using the Amazon Mechanical Turk crowdsourcing platform. Basic demographics were then analyzed to understand the population who uses video resumes to find a job, and results showed that applicants mainly consisted of young people looking for internship and junior positions. We developed a computational framework for the prediction of organizational first impressions, where the inference and nonverbal cue extraction steps were fully automated. Results demonstrate automatic prediction of first impressions of up to 27% of the variance explained for extraversion, and up to 20% for social and communication skills.**

***Index Terms*—Social computing, nonverbal behavior, hirability, personality, video resumes, YouTube, recruitment social videos.**

I. Introduction

SOCIAL media is changing the landscape of personnel recruitment. Beyond the success of LinkedIn and its 300+ million users from 200 countries [43], video interviews are beginning to modify the way in which applicants get hired. Until now, resumes were among the most widely used tools for the screening of job applicants in the personnel selection process [38]. The advent of inexpensive sensors and the success of online video platforms (*e.g.*, YouTube) has enabled the introduction of the video resume. These are short video-recorded messages where job applicants present themselves to potential employers [25]. In comparison with traditional paper resumes, video resumes offer the possibility for applicants to show their personality and communication skills by displaying behavioral information visually and aurally [25], which makes video resumes similar to other forms of online video (*e.g.*, video blogs) in terms of setting [12]. Video resumes constitute an emerging phenomenon. Related work on video resumes is scarce and has focused on their reception by recruiters [25] [30]. To our knowledge, no study has investigated video resumes from a behavioral standpoint. Video resumes hosted on online video platforms represent a unique opportunity to study the formation of first impressions in an employment context at a scale never achieved before.

In our everyday lives, many decisions or judgments people make about others are made based on inferences arising from brief interactions. Social psychology research has shown that humans are accurate at making inferences about others, even if the information is minimal: short displays of behavior can be predictive of social constructs (*e.g.* personality, shyness, or competence) and outcomes (*e.g.* teacher ratings) [6]. In such brief excerpts of social exchanges, nonverbal behavior plays an important role [6]. Nonverbal behavior comprises everything that is transmitted by means other than words and can be perceived aurally (through tone of voice, amount of time spoken, etc.) and visually (through gaze, head gestures, facial expressions, body posture, or hand gestures). It is often the result of unconscious processes, which makes it difficult to fake [31]. Nonverbal behavior has been shown to be a channel through which we reveal our internal states [31] or our personality traits [17].

In this work, we analyze the formation of job-related first impressions in online conversational video resumes. We approach this problem from a nonverbal perspective, where feature extraction and statistical inference are fully automated. We collected a dataset of 939 conversational English-speaking video resumes from YouTube. Annotations of demographics, skills, and first impressions were collected using the Amazon Mechanical Turk (MTurk) crowdsourcing platform [1]. Basic demographics were analyzed to understand the population using video resumes to find a job. The structure of perceived job-related skills was then analyzed in a data-driven clustering experiment. To obtain a feature representation of the video resumes, we extracted nonverbal cues from the audio and visual modalities, hypothesizing that first impressions were at least partly formed based on nonverbal behavior. The linear relationships between nonverbal behavior and the organizational constructs of hirability and personality were examined in a correlation analysis. Finally, we conducted experiments to assess the amount of variance that could be explained by automatically extracted nonverbal features, and analyzed the predictive validity of feature groups. Results showed that most constructs could be inferred significantly more accurately than the baseline-average model, with up to 27% of the variance explained for extraversion, and over 20% for social and communication skills.

The contributions of this work are the following. First, ours is the first study examining conversational video resumes, both from the behavioral and computational perspectives. This setting, which is a new trend in social media and personnel recruitment, has not been studied in computing, and only marginally in management and psychology. Second, we have

collected a dataset of YouTube videos related to personnel selection, which is over 50 times larger than existing related datasets [35], [34]. Third, this study constitutes the first attempt to infer social constructs related of hirability from online videos, and we show that this task can be achieved up to a certain extent. Fourth, we analyze the structure of social skills, and we found that job-related first impressions are mainly formed on one-dimensional positive/negative basis, and that skills can be clustered into three main factors. Fifth, we designed a new instrument to assess job-related skills through the use of crowdsourcing.

Figure **??** displays a summary of this work. This paper is structured as follows. In Section II, we discuss the related work. In Section III, we present the method used for the data collection of video resumes from YouTube. In Section IV, we present and analyze the crowdsourcing experiments designed to collect annotations of demographics, skills, and first impressions of personality and hirability. In Section V, we analyze the structure of perceived skills in a clustering experiment. In Section VI, we present the nonverbal cues extracted from the audio and visual modalities. In Section VII, the linear relationships between nonverbal cues and personality and hirability are investigated in a correlation analysis. In Section VIII, we present the experiments conducted to infer personality and hirability variables using nonverbal cues as predictors. We conclude in Section X.

II. Related work

Despite the emergence of video resumes and video interviewing platforms as a new medium for personnel recruitment, publications on this topic are still scarce. The first references to video resumes date from 1992, when Kelly *et al.* [29] proposed to use video resumes as a tool to support deaf college students develop communication skills to help them secure a job position. Rolls *et al.* proposed in a conceptual study the use of video resumes as a way to "supply the potential employer with insight into the student's personality and character" [39]. Recently, a doctoral thesis by Hiemstra [25] examined the use of video resumes in the personnel selection process. The main focus of this work was to investigate the fairness and discriminatory effects of paper and video resumes, but did not investigate the role of nonverbal behavior in the formation of first impressions. Kemp *et al.* [30] examined the perception of video resumes by sales recruiters. Their study identified general perceptions of video resumes among recruiters and their reactions to this format as a screening tool. To the best of our knowledge, the effect of behavior on first impressions has not been examined in video resumes.

In social media research, Biel *et al.* [12] investigated the formation of personality impressions in conversational video blogs. To this end, a dataset of 442 conversational video blogs was collected from YouTube, and personality impressions were annotated by naïve judges on Amazon Mechanical Turk. To understand the basis on which personality impressions were made, both nonverbal [12] and verbal [14] behavioral features were automatically extracted from the videos and used to infer personality impressions. The studied data was not related to job search as video resumes.

Employment interviews have been studied by organizational psychologists for decades, and one of the main focus is to understand the influence of nonverbal behavior on hirability impressions [17] [27]. The applicant's nonverbal behavior was found to have a remarkable impact: applicants who display more smiles, more eye contact, more facial expressions, and have their body oriented more towards the interviewer have been perceived as more hirable, motivated, competent, and successful than applicants who do not [27] [21]. Most psychology studies on employment interviews relied on manual annotations of nonverbal behavior, which prevents analysis at large scale.

The advent of inexpensive audio and video sensors in combination with improved perceptual methods have enabled the automatic and accurate extraction of behavioral cues, allowing the development of computational methods for inference of individual and group variables such as personality, dominance, or affect [22]. We developed a computational framework for inference of hirability in employment interviews, and demonstrated that predicting hirability impressions using automatically extracted nonverbal cues in combination with machine learning methods was a promising task [35]. Naim *et al.* [34] extended our work by incorporating verbal content and facial expressions, as well as sixteen social traits, such as friendliness, excitement, or engagement. Related to video resumes and job interviews, Batrinca *et al.* [9] used a computational approach to infer Big-Five personality traits in self-presentations, where participants had to introduce themselves in front of a computer in a setting similar to video resumes or video interviews. Similarly, the same authors [8] inferred personality traits in a human-computer collaborative task in a setting similar to video resumes (*i.e.*, a participant facing a computer screen), where subjects had to give instructions to a remote experimenter. However, due to the small size of all previous datasets ($N \in [45, 138]$), the significance and generality of the results were somewhat limited; in this work, we increase the number of subjects by at least one order of magnitude.

We believe that online video resumes constitute a unique setting to study the formation of organizational first impressions at a scale not achieved before. Through Amazon Mechanical Turk, we leverage crowdsourcing mechanisms to obtain inexpensive, fast, reliable, and scalable annotations not only of facts, but also of hirability and personality impressions on online video resumes. We approach the problem of predicting personality and hirability impressions from a computational, nonverbal perspective, where cue extraction and inference are fully automated.

III. Data collection

In this section, we describe the method used to collect a dataset of conversational video resumes from YouTube. Conversational videos are videos where the person mostly speaks in front of a camera. We focused on conversational video resumes because they constitute the setting where behavior can play an important role in the formation of first impressions. Additionally, we concentrated on English-speaking video re-

TABLE I
LIST OF KEYWORDS AND CHANNELS USED TO QUERY YOUTUBE FOR VIDEO IDS.

Keywords:	video curriculum, video cv, video curriculum vitae, vatel cv, video internship cv, video internship resume, digital cv, video resume
Channels:	vatellosangeles, CURRICULUM VIDEO - votre CV VIDEO en studio, Zookel - video Cvs, impressionsD-WREC, Video Resume Now

sumes to ensure that raters understand the verbal content of the videos.

Potential video resumes were queried using the public YouTube Data API [5] to search by keyword and by channels. Keywords were selected by manually browsing through YouTube, and channels found to be specialized in video resumes were also used; Table I displays the keywords and channel IDs used. In total, 5043 unique video IDs were returned by these queries. Videos were downloaded at a maximum of VGA resolution (640×480). The videos were manually filtered to only keep conversational video resumes. To this end, we used a custom-built script to view and keep/discard videos based on keyboard shortcuts. From the 5043 videos downloaded, 1805 videos were rated as conversational video resumes. Crowdourced annotations of language were completed for these videos using Amazon Mechanical Turk [1], among other variables (see Section IV); 939 were rated as English-speaking. These 939 video resumes form the dataset used in the rest of the paper.

This dataset constitutes an increase of the number of subjects by over an order of magnitude, compared to job interview datasets recorded in controlled environments [35] [34]. This increase of N, however, comes at a cost: in comparison with laboratory settings, where high-resolution cameras and professional microphones are used in a quiet environment with optimal illumination, YouTube video resumes feature a wide variety of challenges. The presence of music, improper illumination, low-quality audio, low video resolution, non-fixed cameras, improper framing, text displays and overlays, highly textured background, multiple people on the video, and unexpected user behavior count among the challenges present in online video resumes.

The median duration of the video resumes included in the dataset is $123.5s$, and the distribution is long-tailed ($min_{dur} = 5.1s$, $max_{dur} = 818s$). The median spatial resolution of the videos is 540×360 pixels, while the face resolution is 160×180 pixels in the best-case scenario, *i.e.* when the video is high-resolution and the shot includes the shoulders and face of the subject.

Last, this dataset is considered as personal data by the Swiss Data Protection Law. For this reason, it cannot be distributed. However, we plan to distribute the crowdsourced annotations (Section IV) and the extracted features (Section VI).

IV. CROWDSOURCED ANNOTATIONS OF VIDEO RESUMES

One of the objectives of this work is to understand the demographics of job-seekers uploading video resumes on YouTube, and what types of jobs they apply to. Such statistics on video resumes have not been reported previously. Another objective is to assess the reliability of unacquainted naïve judges in the formation of first impressions on conversational video resumes, including skills and traits. To address these questions, Amazon Mechanical Turk (MTurk) [1] represents an affordable, fast, and fully scalable method to collect annotations of videos. In this section, we present the crowdsourced annotations of video resumes collected using MTurk.

A. Method

Four Human Intelligence Tasks (HITs), *i.e.* individual tasks that MTurk workers work on, were designed to annotate (1) basic facts and demographics (gender, language, audio/video quality); (2) further demographics (age, ethnicity, seniority level, job categories); (3) perceived skills; and (4) first impressions of hirability and personality. In order to ensure that MTurk workers watched a part of the video (*i.e.*, to prevent spammers), the videos had to be watched for a minimal duration (15 seconds for the Basic Facts HIT, 45 seconds for the other HITs) before being able to start answering the questions. For all HITs, all videos were annotated by 5 MTurk raters for a price ranging between \$0.15 and \$0.20 per HIT. To be able to work on the HITs, MTurk workers needed to have over 95% positive feedback in their MTurk previous jobs and had to be located in the U.S.

Crowdsourced annotations given by each rater were aggregated to obtain one score per video and per variable. For categorical variables, the aggregation was obtained by majority voting. Likert variables were considered as continuous (even if strictly speaking this was not the case), and the aggregation was obtained by taking the average value over all judges.

To assess the reliability of each annotated variable in the absence of ground truth, we used inter-rater agreement measures. For categorical variables, (*e.g.* ethnicity, gender), we used Fleiss' Kappa to assess inter-rater reliability. This statistic is similar to Cohen's Kappa, but was designed for any number of raters giving categorical ratings. Fleiss' Kappa assumes a fixed number of raters, but items are not assumed to be rated by the same individuals, which makes this statistic suitable for the task at hand. Fleiss' Kappa can be interpreted as the degree of agreement between the raters above the level of agreement expected by chance [20], but is sensitive to unbalanced categories and depends on the number of categories, therefore its interpretation should be handled with caution. For numerical variables, we used one of the Intraclass Correlation Coefficients (ICC) to assess the level of agreement among judges, as they are commonly used in psychology for this task [42]. $ICC(1,k)$ assesses the degree of agreement in rating the targets (*i.e.*, video resume) when the ratings are aggregated across the judges, and each target is assumed to be rated by a different set of judges, therefore it constitutes the most suitable metric for our task. $ICC(1,k)$ values can be problematic to interpret as no standard threshold exists to segment between *e.g.* moderate and high agreement. To address this, we compared our results with the literature investigating the agreement of judges on related social dimensions. Despite a lack of a clear-cut interpretation for $ICC(1,k)$ values, we used a threshold of $ICC(1,k) = .50$ as a cut-off between low and high inter-rater agreement.

B. HIT 1: Basic facts and demographics

This HIT was used in Section III for the annotation of English-speaking video resumes. In addition to gender (male *vs.* female) and language (Engish *vs.* other), we designed the HIT to verify whether the selected videos were conversational video resumes, using two categorical variables, namely conversational (conversational *vs.* non-conversational *vs.* not sure) and video resume (video resume *vs.* other *vs.* not sure). To this end, we provided MTurk raters with guidelines about what was a conversational video and a video resume, with examples and counter examples. Three questions were asked about the perception of the video quality: video quality, audio quality, and overall quality, which were answered on a 5-point Likert scale. In total, all 1805 videos saved after the manual filtering step (Section III) were annotated.

The level of agreement was high for gender ($\kappa = .97$) and language ($\kappa = .92$). Lower values were found for labelings of conversational videos ($\kappa = .11$) and video resumes ($\kappa = .13$), but this can be explained by the fact that the aggregated class distribution was strongly unbalanced (94.5% *yes* for conversational and 99.5% *yes* for video resumes). Moreover, 60.2% and 92.1% of the videos had full agreement for conversational and video resume, respectively, which shows that agreement on these variables was high despite the low κ values. Finally, the level of agreement for audio, video, and overall quality was high ($ICC(1,k) \in [.75, .77]$). These results suggest that the HITs were conscientiously completed by MTurk workers, and that the manual selection of conversational video resume was consistent with the MTurk annotations.

C. HIT2: Further demographics

This HIT was designed to collect information beyond gender and language, and included the seniority level, age, ethnicity, language proficiency, and the job category of the job seekers depicted in the video resumes. The motivation to collect such annotations was to understand the demographics of job-seekers. Nine types of job categories were used based on the American Time Use Survey (ATUS) [2], a tool developed by the U.S. Bureau of Labor Statistics to measure the amount of time people spend doing various activities, such as paid work, childcare, volunteering, and socializing. ATUS reports statistics of over 140,000 U.S. employees based on occupations, therefore we believe that it constitutes a reliable categorization system for our task. The list and descriptions of the job categories used for this HIT are shown in Table II. The descriptions for each job category were provided in the HIT, and each job category was rated on a 5-point Likert scale. All 939 videos of the video resume dataset were annotated.

The level of agreement among judges was high for age ($ICC(1,k) = .87$), language proficiency ($ICC(1,k) = .89$), ethnicity ($\kappa = .66$), and seniority ($ICC(1,k) = .59$). Table III displays the inter-rater agreement for job categories. These results show that most job categories were reliably annotated with $ICC(1,k)$ values ranging from .58 to .85, with the exception of production ($ICC(1,k) = .31$) and office ($ICC(1,k) = .44$). These results suggest that MTurk raters were reliable in the task of annotating the type of position job-seekers were applying to, and that the annotation of job categories in a crowdsourcing setting is a feasible task.

TABLE II
DESCRIPTIONS OF THE JOB CATEGORIES ANNOTATED IN SECTION IV-C.

Job category	Description
Construction	Construction, installation, maintenance, repair.
Creative	Design, advertisement, music, and arts.
Healthcare	Nurses, doctors, and personal care.
Hospitality	Accommodation, restaurants and bars, travel and tourism.
Management	Management, business, and financial occupations.
Office	Office and administrative support.
Production	Production, transportation, and material moving.
Professional	Computer, engineering, or scientific occupations.
Sales	Sales-related occupations.

TABLE III
INTER-RATER AGREEMENT FOR JOB CATEGORIES, USING $ICC(1,k)$ ($N_{videos} = 939$, $N_{raters} = 5$).

	ICC(1,k)	Mean	STD
Construction	0.62	1.40	0.60
Creative	0.74	2.10	1.02
Healthcare	0.70	1.32	0.57
Hospitality	0.85	2.06	1.18
Management	0.72	2.79	1.06
Office	0.44	2.44	0.82
Production	0.31	1.33	0.43
Professional	0.58	3.19	0.97
Sales	0.63	2.01	0.85

D. HIT3: Skills

This HIT was designed to assess the reliability of naïve judges in the task of rating work-related skills, rather than job domains. Note that by skills we denote an umbrella term including actual skills, traits, and states. We used an initial list of 25 skills known to be often assessed in paper resumes and during job interviews [18]. The list of skills and their descriptions are shown in Table IV. Skills descriptions were provided in the HIT. MTurk raters were asked to answer the question "I see the person as...", and had to rate the person's skill on a 5-point Likert scale. A subset of 200 randomly sampled videos from the video resume dataset was annotated.

TABLE IV
DESCRIPTIONS OF THE SKILLS USED ANNOTATED IN SECTION IV-D.

Skill	Description
Clear	Easy to perceive, understand, or interpret.
Communicative	Willing, eager, or able to talk or impart information.
Competent	Having the necessary ability to do something successfully.
Concise	Giving a lot of information clearly in a few words.
Confident	Feeling/showing confidence in oneself or one's abilities.
Creative	Involving the use of original ideas to create something.
Dedicated	Devoted to a task or purpose.
Empathic	Able to understand and share the feelings of others.
Enthusiastic	Having/showing intense and eager enjoyment or interest.
Friendly	Kind and pleasant.
Funny	Causing laughter or amusement; humorous.
Hard-Working	Tending to work with energy and commitment; diligent.
Honest	Free of deceit; truthful and sincere.
Independent	Not relying on others for aid and support.
Intelligent	Having or showing intelligence, especially of a high level.
Leader	A person who has the ability to lead a group/organization.
Motivated	Enthusiastic and determined to achieve success.
Open-Minded	Willing to consider new ideas; unprejudiced.
Organized	Able to plan one's activities efficiently.
Persuasive	Good at persuading someone to do or believe something.
Positive	Constructive, optimistic, or confident.
Professional	Worthy of a professional person; competent, skillful, or assured.
Reliable	Consistently good in quality or performance; able to be trusted.
Sociable	Willing to talk and engage in activities with other people.
Stressed	Mentally tensed and worried.

TABLE V
INTER-RATER AGREEMENT FOR PERCEIVED SKILLS, USING $ICC(1,k)$ ($N_{videos} = 192$, $N_{raters} = 5$). VARIABLES WITH AN ASTERISK* WERE NOT USED IN SECTION V DUE TO THEIR LOW $ICC(1,k)$ VALUES.

	ICC(1,k)	Mean	STD
Clear	0.74	3.50	0.82
Communicative	0.68	3.84	0.67
Competent	0.62	3.99	0.63
Concise	0.59	3.56	0.71
Confident	0.67	3.88	0.69
Creative	0.65	3.35	0.69
Dedicated	0.51	4.00	0.55
Empathic*	0.49	3.33	0.52
Enthusiastic	0.68	3.64	0.72
Friendly	0.63	3.91	0.59
Funny	0.55	2.60	0.68
Hard-Working	0.55	4.04	0.56
Honest*	0.36	4.02	0.44
Independent	0.55	4.00	0.57
Intelligent	0.52	4.02	0.55
Leader	0.62	3.32	0.72
Motivated	0.57	4.06	0.57
Open-Minded*	0.38	3.73	0.48
Organized	0.63	3.81	0.62
Persuasive	0.67	3.33	0.71
Positive	0.61	3.99	0.56
Professional	0.70	3.86	0.75
Reliable	0.60	4.00	0.56
Sociable	0.65	3.70	0.65
Stressed*	0.25	2.05	0.54

For these perceived skills, intraclass correlation coefficients are shown in Table V. Generally, MTurk workers were reliable in rating perceived skills, with 21 out of 25 skills with $ICC(1,k) > .50$. Some of the skills (persuasive, creative, professional, clear, confident, enthusiastic) had reliabilities greater than .65, which is comparable to the ones obtained for job categories. Honest, open-minded, empathic, and stressed were observed to have low inter-rater agreement ($ICC(1,k) < .50$), suggesting that these skills were not evident to rate, or that the setting did not elicit clear displays of such skills. The low inter-rater agreement obtained for stressed can be explained by the fact that a high value indicated a negative perception, which was the opposite for all other skills; this could have confused the raters. Additionally, stress has been previously reported as a variable with relatively low inter-rater agreement in social video [41].

E. HIT4: First impressions

In this HIT, we asked MTurk raters to give their first impressions on a video resume. Specifically, MTurkers were asked to rate two variables for the general first impression (overall hirability and overall first impression), three high-level skills derived from the clustering of skills (professional, social, and communication skills, see Section V for details), and ten items for perceived personality. For personality, we used the Big-Five model which represents personality at its highest level of abstraction, and which consists of five factors, namely extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness [23]. To assess personality, we used the standard Ten Item Personality Inventory (TIPI) questionnaire consisting of ten items, two per dimension [23]. Each question was answered on a five-point Likert scale. All 939 videos of the video resume dataset were annotated.

Table VI displays the inter-rater agreement for the first impressions HIT. Hirability first impressions (overall hirability, overall impression, and professional, social, and communication skills) were observed to have moderate-to-high inter-rater agreement with $ICC(1,k) \in [.59, .64]$. These values are comparable to the ones obtained for low-level skills, and somewhat lower than the job categories, but still demonstrate that reliable first impressions of hirability from naïve raters can be obtained from video resumes. For personality, only extraversion was observed to be consistently rated ($ICC(1,k) = .64$). Openness to experience showed moderate-to-low agreement ($ICC(1,k) = .46$), while agreeableness, conscientiousness, and neuroticism had low inter-rater agreement ($ICC(1,k) \in [.18, .38]$). In comparison with other works investigating the reliability of personality impressions, several points can be noted. First, extraversion was the trait that achieved the highest level of agreement, which has been repeatedly observed in related work [12], [24]. Second, the overall agreement on personality impressions was lower than the one obtained from video blogs [12]. In order to understand whether this relatively low inter-rater agreement was due to how the first impression HIT was designed or to the setting itself, we collected a second round of MTurk annotations on a subset of the data ($N = 200$), where only personality traits were rated using the TIPI questionnaire [23]. Inter-rater agreement results resulting from this experiment were observed to be very close to the ones reported in Table VI and are not shown here in detail. This suggests that with the exception of extraversion (and openness to experience, to a lesser extent), personality impressions were difficult to rate consistently from video resumes. This could be explained by the fact that personality traits other than extraversion and openness to experience might not be expressed in this setting. This hypothesis would have to be validated as part of future work.

TABLE VI
INTER-RATER AGREEMENT FOR ANNOTATIONS FROM THE FIRST IMPRESSION HIT, USING $ICC(1,k)$ ($N_{videos} = 939$, $N_{raters} = 5$).

	ICC(1,k)	Mean	STD
Overall Impression	0.59	3.70	0.62
Overall Hirability	0.61	3.72	0.62
Professional Skills	0.59	3.76	0.60
Social Skills	0.57	3.67	0.63
Communication Skills	0.64	3.71	0.69
Extraversion	0.64	3.46	0.61
Agreeableness	0.27	3.71	0.39
Conscientiousness	0.38	3.91	0.46
Neuroticism	0.18	2.21	0.40
Openness	0.46	3.51	0.51

F. Demographics of video resumes

Figure 1 displays the demographics (gender, age, ethnicity, job type, seniority level, and duration) of the video resume dataset. We first observe that the majority of the population was young (59% were under 25 and 93% were under 35) and either looking for an internship position (37%) or a junior position (41%). Gender was unbalanced as there were approximately twice as many males (66%) as females (34%). In terms of ethnicity, the video resume dataset was mainly populated by Caucasian (36%) and Indian (32%) jobseekers. One possible hypothesis to explain this finding stems from the bias generated by the decision of keeping only

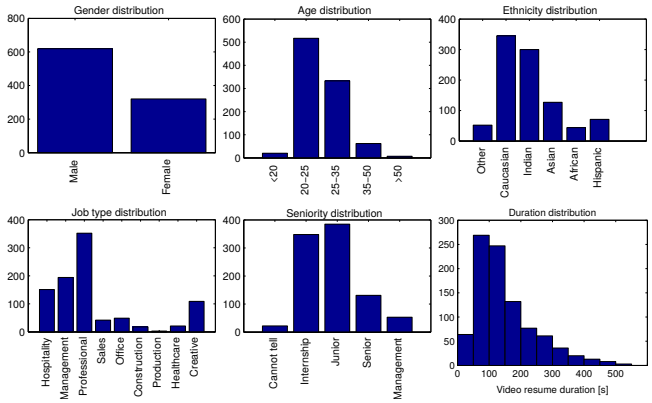


Fig. 1. Demographics (gender, age, ethnicity, job type, seniority, and video resume duration) of the video resume dataset ($N = 939$).

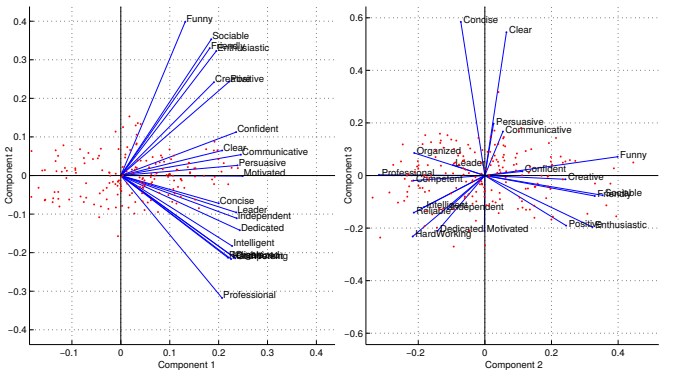


Fig. 2. First three principal components of the principal component analysis (PCA) on perceived skills ($N = 200$), accounting for 82% of the variance. Red dots and blue lines are the data points and the original axes projected onto the PCA space, respectively.

English-speaking video resumes. For instance, many Spanish-speaking video resumes were present in the set prior to filtering out non-English video resumes. For job categories, the most represented job category was professional (computer, engineering, or science occupations) accounting for 39% of the dataset, followed by management (21%), hospitality (16%), and creative (12%). In terms of duration, the majority of the video resumes are relatively short, with 62% lasting less than 150 seconds, and the main mode situated between 50 and 100 seconds. The distribution was long-tail like, and the maximum video resume duration in the dataset is 818 seconds.

Demographics varied across job categories. The professional category (engineers, scientists, and computer scientists) was populated by 77% males, 50% Indians, and 26% Caucasians. One missing part in our dataset is the geographic distribution of uploads. this could explain possible biases related to the home location of the applicants. For hospitality and creative applicants, the gender was more balanced (48% and 57% males, respectively) and these categories were dominated by Caucasian applicants (28% and 64%, respectively). In terms of age and seniority distributions, applicants in the management category were older and were applying for more senior positions, and no major difference between the professional, hospitality, and creative job categories was observed.

V. Clustering of Skills

To understand the structure of the 25 perceived job-related skills (listed in Table IV), we conducted a clustering analysis of the skills annotated in the Skills HIT (Section IV-D). The values were first pre-processed: skills with low inter-rater agreement (($ICC(1,k) < 0.5$) were removed, and each variable was standardized such that it had zero mean and unity variance. This left a set of 21 skills.

We then conducted principal component analysis (PCA) on the skills variables. Figure 2 displays the video resume data points as well as the original variables projected onto the coordinate system of the three first principal components, accounting for 82% of the variance in the data. One can observe that the first component, accounting for 64% of the variance, was not very distinctive of skills as they were all positive, but was evocative of the fact that video resumes were rated mainly on a good/bad basis. In psychology, this effect is known as the halo effect, where global evaluations tend to induce altered evaluations of other attributes [36]. The second principal component, accounting for 14% of the variance, seemed to encompass two different social concepts. On one hand, the variables of funny, sociable, friendly, enthusiastic, creative, positive seemed to be part of a *social* high-level label; on the other hand, *professional* variables seemed to form a cluster including competent, organized, professional, reliable, hard-working, etc. The third component accounted for 4% of the variance, and seemed to enclose *communication* skills: concise, clear, persuasive, and communicative.

K-means clustering of the original variables was performed in the principal component space. We used the Euclidean distance as distance measure, but PCA coordinates were multiplied by the square-root of the eigenvalues corresponding to each component (thus accounting for the standard deviation explained by each principal component) prior to completing K-means clustering. Experiments with $K \in [2, 7]$ were completed, and $K = 3$ appeared to be a subjectively optimal choice.

Figure 3 displays the three obtained clusters of perceived skills, where colors denote correlation coefficients (warm is positive, cold is negative). A first dense cluster including organized, motivated, independent, dedicated, intelligent, competent, professional, hard-working, leader, and reliable was observed, and was labeled as *professional* skills. The second cluster encompassed creative, friendly, enthusiastic, positive, funny, and sociable, and was labeled as *social* skills. The last cluster included persuasive, clear, concise, communicative, and confident, and was labeled as *communication* skills. As a verification step, factor analysis was also completed on the skills, and similar clusters were found.

The three obtained high-level skills of this Section were then annotated (as already discussed in Section IV-E) for all video resumes and used as hirability variable for the correlation and inference analyses (Sections VII and VIII).

VI. Extraction of Behavioral Cues

One of the objectives of this work is to investigate the use of automatically extracted behavioral cues for the inference of first impressions in conversational video resumes. To this end, we used existing methods to extract features from both

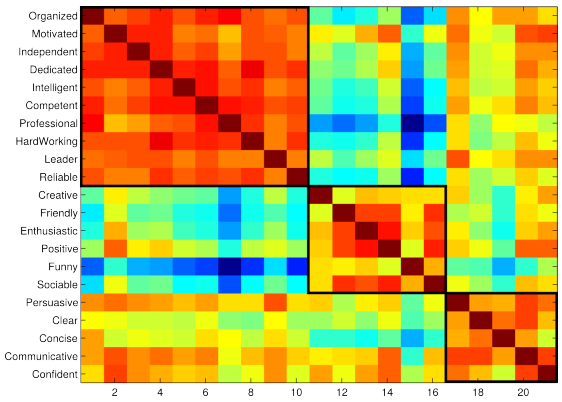


Fig. 3. K-means clustering on skills ($N = 200$), with $K = 3$. Warm and cold colors denote positive and negative correlation coefficients, respectively.

the audio and visual modalities. The choice of nonverbal cues to be extracted was based on the nonverbal communication literature [31] and the existence of available processing methods. Because job-seekers do not present themselves in front of another person, conversational video resumes cannot be considered as face-to-face social interaction as such. Nevertheless, they consist of a person delivering a message in a natural manner, and we hypothesize that the findings in social psychology stating that nonverbal behavior influences the way people are perceived still apply [31]. Recent related work in social computing has observed that nonverbal cues can be predictive of social constructs in non-face-to-face interactions such as video blogs [12] or human-computer interfaces [9].

In comparison with datasets recorded in laboratory settings using high-end audio and video sensors and optimal illumination, user-generated online videos are challenging to process. The presence of music, improper illumination, low-quality audio recording, high compression rates, low video resolution, non-fixed cameras, text displays and overlays, multiple people on the video, and unexpected user behavior count among the many challenges for automatically processing online video resumes. As a result, simple and robust extraction methods were preferred over fine-grain but more error-prone algorithms. Furthermore, the set of extracted behavioral features was assumed to be inherently noisy to the challenging nature of online video resumes.

Please note that in this paper, we use the terms ”behavioral cues”, ”nonverbal cues” and ”features” interchangeably to designate the numerical representation of video resumes.

A. Audio cues

Audio nonverbal cues are associated with a broad array of social constructs in social psychology [31], and have been successfully used in computational applications for the automated inference of constructs as diverse as dominance [28], emergent leadership [40], personality traits [37], negotiations outcomes [16], or hirability ratings [35] in face-to-face settings. Interestingly, audio nonverbal cues have also been successfully used for the inference of personality traits in settings similar to video resumes, such as in human-computer interfaces [9] or video blogs [12], even if no conversational partner was included in these scenarios.

Similarly to most the related works, we extracted nonverbal cues based on speaking activity and prosody. To this end, we used the MIT Media Lab Speech Extraction Code [4]. This open-source software package consists of a two-step hidden Markov model (HMM) used to segment the speech signal into speech/non-speech and voiced/non-voiced regions at a frame rate of 62.5 Hz [7]. As a second step, the fundamental frequency of the audio signal is tracked over voiced regions, using a probabilistic method [7]. No validation of the speech activity extraction step was performed, but we computed the histogram of the detected speech activity ratio (*i.e.*, the ratio between the number of frames labeled as speech and the total number of frames, displayed in Figure 4), which shows that the majority of video resumes had over 50% of the frames labeled as speech.

1) Speaking activity: We defined a speaking activity event as a sequence of audio frames where speech was detected. Statistics on the speaking activity event durations were computed for each video resume: mean, median, standard deviation, minimum, maximum, and quartiles, as well as the number of turns and the ratio of speaking time.

2) Prosody: Prosody describes all the variations that accompany speech in the vocal delivery, and is composed of three fundamental acoustic properties, namely speech rate (the number of voiced segments per unit of time), pitch (the fundamental frequency), and intensity (the energy of the speech sound, perceived as loudness) [31].

- **Voiced rate**. Cues related to speech rate were derived from the voiced/non-voiced segmentations obtained from the Speech Feature Extraction Code [4]. Voiced events were defined as a sequence of audio frames classified as *voiced*, and were characterized by their starting and ending times. Statistics on the durations of voiced segments were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy, as well as the ratio of voiced time over speaking time.
- **Speech energy**. Cues related to speech energy were obtained by squaring the value of each audio sample classified as *speaking*. Statistics over the time-series were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.
- **Pitch**. Cues related to pitch were derived from the pitch values obtained from the Speech Feature Extraction Code [4]. Statistics of the pitch values were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.

B. Visual cues

The visual modality of the video resumes was characterized by nonverbal cues capturing proximity, frontal face events, head motion, and body activity. These visual cues were successfully used to characterize people in a setting similar to video resumes, namely video blogs [12], and have the main advantage to be easy to extract.

1) Proximity, frontal face events, and head motion: Our aim is to capture the proximity, looking behavior, and head motion of the video job-seekers, but due to the lack of accurate extraction methods, we used frontal face detections as a surrogate.

To this end, we used a standard frontal face detection system based on the object detection method proposed by Viola and Jones [45] for the extraction of proximity, frontal face events, and head motion cues. Over more elaborate tracking algorithms, this method has the advantages of being robust, simple to implement, and not requiring any initialization or parameter tuning. For each frame, the algorithm extracted the bounding box of each detected frontal face, characterized by its position and size. The frontal face detection method was not systematically evaluated, but the output of the algorithm was manually inspected for a subset of 100 randomly sampled video resumes, and the faces were overall correctly detected in over 85 videos. Also, we computed the histogram of the ratio between the number of detected frontal faces and the total number of frames (see Figure 4), and the large majority of video resumes had an important ratio of detected faces, which is an indicator of the correct detection of faces in the context of conversational videos. Among the observed errors, intermittent detection within a video, or false positives detected for a short duration were the most frequent. In some rare cases, no face could be detected for the whole duration of the video, which was generally due to improper framing and excessively bad illumination and/or overall video quality. To account for multiple face detections due *e.g.* to the presence of pictures including faces or other objects with face-like appearance, we hypothesized that the job seeker's face corresponded to the largest and highest bounding box, and discarded all other detected faces based on a simple cost function. Intermittent detections were also handled by removing each sequence of detected faces that lasted less than half a second, and by assigning sequences of non-detections shorter than half a second to the closest detected bounding box.

- **Proximity**. We used the size of the detected face bounding boxes as a proxy for the distance between the person's face and the camera. To account for the variability of video resolution across video resumes, the size of the bounding boxes was divided by the overall resolution. Statistics over the size of the bounding boxes were computed: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.
- **Head motion**. As a proxy for head motion, we used the displacement of the detected bounding box centers of two consecutive frames, and computed statistics for the horizontal and vertical displacements, as well as for the magnitude: mean, median, standard deviation, minimum, maximum, quartiles, and entropy.
- **Frontal face events**. We used detected frontal faces as a proxy for looking at the camera. We defined frontal face events as a sequence of frames where a frontal face was detected, and each event was characterized by its starting and ending time. To encode frontal face activity patterns, we computed statistics on the duration of frontal face events: mean, median, standard deviation, minimum, maximum, and quartiles, as well as the ratio between the number of frontal face frames and the total number of frames.

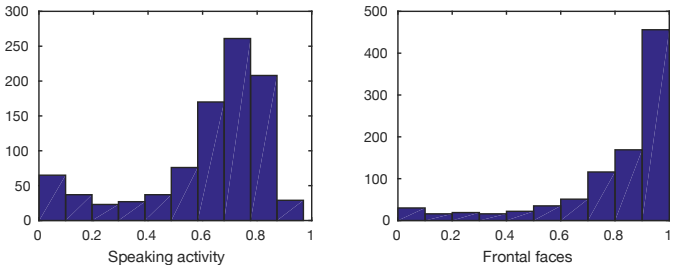


Fig. 4. Histograms of detected speech activity (left) and detected frontal face ratios (right).

2) Overall visual motion: The overall visual motion of a job-seeker in a video resume can indicate his level of kinetic expressiveness. To measure the overall movement, we used the Weighted Motion Energy Image (WMEI) descriptor [10], which is a modified version of motion energy images. WMEIs summarize the motion throughout a video as a single grayscale image, where each pixel intensity indicates the visual activity at its position. As descriptors for the overall visual motion, we computed statistics on the WMEIs: mean, median, standard deviation, minimum, maximum, entropy, quartiles, and center of gravity.

C. Video statistics

In order to obtain indicators of the overall quality of video resumes, we extracted basic video statistics using ffmpeg [3]. These statistics include the size, duration, bitrate, audio sampling rate, horizontal, vertical, and overall resolution, and the framerate. Although not behavioral as such, these features have the main advantage of being simple to extract, and can be seen as an indicator of the effort put by job-seekers in the process of recording their video resumes.

D. Feature pre-processing

Prior to carrying out further analyses, nonverbal cues were pre-processed. Data for which no face or no speech was detected were discarded. In total, 882 videos were kept for analysis (57 removed). Highly skewed features ($skew > 1$) were log-transformed ($x' = log(1+x)$ where x' and x denote the transformed and original features, respectively). Finally, all nonverbal cues were standardized, such that each cue had zero mean and unity variance.

VII. Correlation Analysis

A. Personality and hirability

Table VII displays the pairwise correlations between the Big-Five personality variables and the hirability variables. High correlation ($r = .69$) was observed among the personality traits of extraversion and openness to experience, suggesting that job seekers who were rated high on the trait of extraversion were likely to get high scores for openness to experience. The neuroticism trait was observed to be negatively correlated with all other variables, particularly with conscientiousness ($r = -.65$). Other correlations among personality impressions were lower, with $|r| \in [.25, .45]$.

TABLE VII
PAIRWISE CORRELATIONS BETWEEN THE PERSONALITY AND HIRABILITY VARIABLES, USING PEARSON'S CORRELATION COEFFICIENT. ALL VALUES ARE STATISTICALLY SIGNIFICANT ($p < 10^{-12}$), $N = 939$.

	2	3	4	5	6	7	8	9	10
1. Extra.	0.27	0.25	-0.33	0.69	0.55	0.51	0.31	0.66	0.62
2. Agree.		0.39	-0.43	0.43	0.35	0.35	0.24	0.40	0.32
3. Consc.			-0.65	0.40	0.58	0.63	0.62	0.46	0.45
4. Neuro.				-0.45	-0.49	-0.51	-0.45	-0.47	-0.46
5. Open.					0.58	0.54	0.37	0.66	0.63
6. OvImpr.						0.87	0.75	0.79	0.81
7. OvHira.							0.80	0.77	0.78
8. ProSk.								0.59	0.63
9. SocSk.									0.84
10. ComSk.									

Hirability variables were found to be strongly inter-correlated ($r \in [.59, .81]$), and the strongest relationship was observed between overall impression and overall hirability ($r = .87$). Social and communicative skills were also strongly correlated ($r = .84$), and interestingly, were also highly correlated with the personality traits of extraversion and openness to experience ($r \in [.62, .66]$). As for the variable of professional skills, it was associated to the personality variable of conscientiousness ($r = .62$).

B. Nonverbal behavior and personality

Table VIII displays Pearson's pairwise correlation coefficients between the nonverbal cues extracted in Section VI and crowdsourced personality impressions, where only cues with $p < 10^{-3}$ are shown. We first observe that the traits of extraversion and openness to experience were the only personality variables with a large number of low-to-moderate correlated cues. Interestingly, extraversion and openness were also the two traits with acceptable inter-rater agreement ($ICC_{extra}(1,k) = .64$ and $ICC_{open}(1,k) = .46$, see Table VI). Given the low inter-rater agreement of agreeableness, conscientiousness, and neuroticism, it is not surprising to observe low correlation values. The sets of nonverbal cues correlated with extraversion and openness to experience were observed to be very similar and have comparable correlation coefficients. This finding can be explained by the fact that these two traits were highly correlated ($r = .69$, see Table VII).

Surprisingly, only a few vocal cues were correlated with extraversion and openness to experience. Apparently, this finding contradicts previous studies where vocal cues were found to be predictive of personality traits in general, and extraversion in particular, independently of the type of setting [12] [32]. Furthermore, the literature on nonverbal communication has established that extraverted individuals are mainly characterized by a large amount of speech, voice modulation, and energy [31]. One possible hypothesis to explain this finding could be that the errors in the vocal cue extraction process generated noisy features; this however needs to be investigated in future work.

Frontal face events were observed to be correlated with extraversion and openness to experience. In particular, the correlation coefficients for the number of frontal face events was positive, and they were negative for frontal face event durations, which was also observed in video blogs [12].

TABLE VIII
NONVERBAL CUES SIGNIFICANTLY CORRELATED WITH AT LEAST ONE PERSONALITY VARIABLE, USING PEARSON'S CORRELATION COEFFICIENT ($p < 10^{-3}$, $^{*}p < 10^{-4}$, $^{\dagger}p < 10^{-5}$).

NVB cue	Extra.	Agree.	Consc.	Neuro.	Open.
Vocal cues:					
Spk ratio		0.12	0.13^{*}		
STD of spk turn dur.					0.11
Med. spk turn dur.			0.11	-0.11	
Max. spk turn dur.	0.13^{*}				$0.16^{\dagger}$
STD of pitch		-0.12			
Entr. of pitch	$0.15^{\dagger}$				0.14^{*}
Max. voiced dur.	0.12				0.11
Frontal face events (FFE):					
# of FFE	$0.16^{\dagger}$				$0.18^{\dagger}$
Mean FFE dur.	$-0.20^{\dagger}$				$-0.19^{\dagger}$
Med. FFE dur.	$-0.20^{\dagger}$				$-0.19^{\dagger}$
STD of FFE dur.	0.13				
Min. FFE dur.	$-0.21^{\dagger}$				$-0.20^{\dagger}$
Max. FFE dur.	$-0.16^{\dagger}$				$-0.15^{\dagger}$
FFE dur. Q25	$-0.21^{\dagger}$				$-0.20^{\dagger}$
FFE dur. Q75	$-0.18^{\dagger}$				$-0.17^{\dagger}$
Frontal face ratio	-0.14^{*}				$-0.17^{\dagger}$
Proximity cues:					
Mean bbox size	-0.12				
STD of bbox size	$0.16^{\dagger}$				$0.20^{\dagger}$
Med. bbox size	-0.13				
Min. bbox size	$-0.23^{\dagger}$				$-0.18^{\dagger}$
Bbox size Q25	$-0.15^{\dagger}$				
Head motion (HM):					
Mean h. HM	$0.25^{\dagger}$				$0.23^{\dagger}$
STD of h. HM	$0.24^{\dagger}$				$0.24^{\dagger}$
Med. h. HM	$0.18^{\dagger}$				$0.15^{\dagger}$
Max. h. HM	$0.29^{\dagger}$				$0.30^{\dagger}$
H. HM Q25	$0.21^{\dagger}$				$0.16^{\dagger}$
H. HM Q75	$0.19^{\dagger}$				$0.19^{\dagger}$
H. HM entropy	-0.13				-0.13^{*}
Mean v. HM	$0.20^{\dagger}$				$0.22^{\dagger}$
STD of v. HM	$0.18^{\dagger}$				$0.20^{\dagger}$
Med. v. HM	$0.16^{\dagger}$				$0.16^{\dagger}$
Max. v. HM	$0.25^{\dagger}$				$0.26^{\dagger}$
V. HM Q25	$0.18^{\dagger}$				$0.15^{\dagger}$
V. HM Q75	$0.18^{\dagger}$				$0.20^{\dagger}$
Mean HM	$0.25^{\dagger}$				$0.25^{\dagger}$
STD of HM	$0.24^{\dagger}$				$0.25^{\dagger}$
Med. HM	$0.20^{\dagger}$				$0.19^{\dagger}$
Max. HM	$0.30^{\dagger}$				$0.31^{\dagger}$
HM Q75	$0.21^{\dagger}$				$0.21^{\dagger}$
HM entropy	-0.13^{*}				-0.14^{*}
Overall motion:					
Mean WMEI	$0.20^{\dagger}$				$0.19^{\dagger}$
Med. WMEI	$0.20^{\dagger}$				$0.18^{\dagger}$
Min. WMEI	$0.15^{\dagger}$				$0.17^{\dagger}$
WMEI Q25	$0.22^{\dagger}$				$0.23^{\dagger}$
WMEI Q75	$0.19^{\dagger}$				$0.17^{\dagger}$
WMEI entropy	$0.25^{\dagger}$				$0.20^{\dagger}$
Video stats:					
Video size					0.13^{*}
Bitrate	$0.18^{\dagger}$				$0.17^{\dagger}$

This could suggest that job-seekers who did not consistently keep their face fronting the camera were perceived as more extraverted and more open. Under the assumption that frontal faces are a proxy for looking at the camera, this finding would contradict the nonverbal behavior literature that established that gaze was associated with extraversion [31], but frontal face turn breaks could also result from head movements, inserts of short non-conversational video snippets (*e.g.*, the job-seeker showing himself doing something), or detection failures, instead of gaze avoidance. In any case, these results need to be investigated in greater details.

To a lesser degree, proximity cues were also correlated with extraversion and openness to experience. Large variations of face proximity and smaller face bounding boxes were associated with higher ratings on extraversion and openness to experience. These observations differ from but do not contradict previous works, where proximity features were not significantly correlated with these traits in video blogs [12].

TABLE IX
NONVERBAL CUES SIGNIFICANTLY CORRELATED WITH AT LEAST ONE HIRABILITY VARIABLE, USING PEARSON'S CORRELATION COEFFICIENT ($p < 10^{-3}$, $^{*}p < 10^{-4}$, $^{\dagger}p < 10^{-5}$).

NVB cue	OvImpr.	OvHira.	ProSk.	SociSk.	CommSk.
Vocal cues:					
Spk ratio					0.12
Mean energy	-0.11			-0.13	$-0.17^{\dagger}$
STD of energy	-0.11			-0.13	$-0.17^{\dagger}$
Med. energy				-0.11	-0.14^{*}
Max. energy					-0.12
Energy Q25					-0.13^{*}
Energy Q75	-0.11			-0.13	$-0.16^{\dagger}$
Frontal face events (FFE):					
Mean FFE dur.				-0.13^{*}	-0.13
Med. FFE dur.				-0.14^{*}	-0.13^{*}
Min. FFE dur.				-0.14^{*}	-0.14^{*}
Max. FFE dur.				-0.11	
FFE dur. Q25				-0.14^{*}	-0.14^{*}
FFE dur. Q75				-0.12	-0.11
Proximity cues:					
Mean bbox size	-0.11				
Med. bbox size	-0.11				
Min. bbox size	$-0.16^{\dagger}$	-0.12	-0.11	$-0.15^{\dagger}$	-0.11
Bbox size Q25	-0.12				
Head motion (HM):					
Mean h. HM	0.11			$0.16^{\dagger}$	$0.16^{\dagger}$
STD of h. HM	0.13^{*}			$0.16^{\dagger}$	$0.16^{\dagger}$
Med. h. HM	0.12			0.14^{*}	0.13^{*}
Max. h. HM	$0.19^{\dagger}$	$0.15^{\dagger}$	0.12	$0.21^{\dagger}$	$0.20^{\dagger}$
H. HM Q25	$0.18^{\dagger}$	$0.16^{\dagger}$		$0.19^{\dagger}$	$0.20^{\dagger}$
H. HM Q75				0.13	0.12
Mean v. HM	0.11			0.14^{*}	0.14^{*}
STD of v. HM				0.12	0.11
Med. v. HM	$0.15^{\dagger}$	0.12		0.14^{*}	0.14^{*}
Max. v. HM	0.14^{*}			$0.17^{\dagger}$	$0.16^{\dagger}$
V. HM Q25	$0.16^{\dagger}$	0.14^{*}		$0.17^{\dagger}$	$0.16^{\dagger}$
V. HM Q75	0.14^{*}	0.12		$0.16^{\dagger}$	$0.15^{\dagger}$
Mean HM	0.12			$0.17^{\dagger}$	$0.16^{\dagger}$
STD of HM	0.13^{*}			$0.16^{\dagger}$	$0.16^{\dagger}$
Med. HM	0.14^{*}			$0.15^{\dagger}$	$0.15^{\dagger}$
Max. HM	$0.19^{\dagger}$	$0.15^{\dagger}$	0.12	$0.21^{\dagger}$	$0.20^{\dagger}$
HM Q25	$0.15^{\dagger}$	0.12	0.11		
HM Q75				0.14^{*}	0.13^{*}
Overall motion:					
Mean WMEI				0.13^{*}	
Med. WMEI				0.12	
Min. WMEI				0.11	
Max. WMEI			0.11		
WMEI Q25				$0.15^{\dagger}$	0.11
WMEI entropy				0.11	
Video stats:					
Bitrate	0.12	0.12		$0.17^{\dagger}$	$0.17^{\dagger}$

This suggests that relative positioning of the candidate has a small yet significant effect. For head motion, a large number of nonverbal cues were significantly correlated with extraversion and openness, with correlation coefficients up to $r = .30$. The linear relationship between head motion cues and the personality dimensions was mainly positive, meaning that job-seekers who displayed larger head motions were perceived as more extraverted and more open. Although head motion cues were not investigated in previous studies focusing on online videos, this observation was confirmed by the nonverbal communication literature, which has shown that extraversion was associated with kinetic expressiveness, including head movements [31]; the relationship between head motion and openness to experience was however not clearly established in the literature [31] but can be explained by the strong correlation observed between extraversion and openness. For overall body motion measured by statistics on weighted motion energy images (WMEI), cues were also found to be correlated with extraversion and openness. Specifically, job-seekers displaying a larger and more diverse visual activity were perceived as more extraverted and open. Again, the nonverbal communication literature confirms this observation, as extraverted people are usually kinetically expressive [31]; furthermore, similar results were reported on video blogs [12]. Last, some basic video statistics (bitrate and video size) were positively correlated with extraversion and openness to experience, suggesting that high definition video resumes earned higher ratings for these two traits.

C. Nonverbal behavior and hirability

Table IX displays Pearson's pairwise correlation coefficients between the nonverbal cues extracted in Section VI and the hirability variables, where only cues with $p < 10^{-3}$ are displayed. Overall, low to moderate yet statistically significant effects can be observed. First, no vocal cue was correlated with overall hirability, and only three vocal cues were correlated with the overall impression. This is surprising, considering results previously reported on employment interviews [35], where hirability was observed to be correlated to applicant speaking turn-based and prosodic cues. To understand these differences, one should consider the differences of settings between employment interviews and video resumes. First, job interviews recorded in laboratory settings represent an ideal setting for the clean extraction of vocal cues due to the absence of background noise and the use of microphone arrays to segment speaker turns. In video resumes, the audio quality was not guaranteed (background noise, presence of music, low-quality microphones) and the extraction process might have suffered from it. Second, the structured nature of the job interview dataset presented in [35] allowed to objectively compare the behavior of job applicants, which was particularly important for speaking turn-based cues, whereas in video resumes no such structure existed. Cues related to head motion were found to be positively correlated with the overall impression and, to a lesser degree, with overall hirability. This suggests that job-seekers who displayed kinetic expression from the head were positively rated. For overall motion, no WMEI cue was correlated with either impression or hirability. Similar results were observed in employment interviews [35]. For basic video statistics, bitrate was positively correlated with all hirability variables except for professional skills, which indicates that videos of high definition made a better impression on raters.

For social and communication skills, the sets of correlated nonverbal cues were very similar and the correlation values very close. The high correlation found between the two variables ($r = .84$) explains these similarities. Additionally, social and communication skills shared a similar set of correlated cues with the personality variables of extraversion and openness to experience, which can also be explained by the high correlations among these variables ($r \in [.62, .66]$). Compared to other variables, communication skills was correlated to a relatively large number of vocal cues. Specifically, job-seekers who spoke much, and with low energy were perceived as more communicative, and similar observations can be made for social skills to a lesser degree. These observations can be justified by the fact that these skills inherently require speech abilities. The variable of professional skills had a low number of correlated cues, which can either be explained by the hypothesis that the extracted nonverbal cues were too noisy

to hold any relevant information, or that professional skills were rated on the basis of other behavioral cues, *e.g.* verbal cues; this result needs to be investigated in greater detail in the future.

VIII. Inference

One of the objectives of this work is to investigate the use of automatically extracted nonverbal cues for the prediction of personality and hirability impressions. We defined the inference task as a regression task, where the goal was to infer the exact scores of the personality and hirability variables collected in the first impression HIT (see Section IV-E). Two experiments were completed: first, we assessed the performance of several regression and dimensionality reduction methods (Section VIII-A); second, we analyzed the predictive validity of feature groups (Section VIII-B). We used a 10-fold cross-validation approach for training and testing the regression models. This framework used 90% of the video resumes for training, and kept the 10% remaining for testing. Model parameters were estimated using a 10-fold inner cross-validation approach. To quantify the performance of the automatic inference models, we used the root-mean-square error ($RMSE$) and the coefficient of determination (R^2), as these are two widely used measures in the psychology and social computing. As the baseline regression model, we took the average score as the predicted value.

A. Comparison of regression methods

Several standard dimensionality reduction techniques were tested. *Low p-value features (pval)* assumed a linear relationship between features and variables, and that the relevant information is contained in the features significantly correlated with the variable of interest; we selected cues with $p < .05$. *Principal component analysis (PCA)* projects the data onto an orthogonal space of lower dimension; the number of components was set such that 99.9% of the variance was explained by the model. *All features (all)* is the simplest way to test the improvement of the dimensionality reduction step: we tested the case of taking all features as predictors for the regression step.

For regression, we used the following regression techniques. *Ridge regression (ridge)* minimizes the sum of squared errors between the observed and predicted responses of a linear model, and a regularization term is added to the cost function, which multiplies the l_2-norm of the regression coefficients; this regression penalty has the effect of shrinking the estimates towards zero, preventing the model to over-fit. *LASSO regression (LASSO)* minimizes the same cost function than ridge regression, but the regularization term multiplies the l_1-norm of the regression coefficients, resulting in sparse coefficients and preventing the model to over-fit [44]. *Random forest (RF)* is based on the bootstrap aggregation of a large number of decision trees that split the feature space into hyper-cubic regions assigned to values [15]; RF aggregates the output of each separate decision tree by taking the average predicted value and has the advantage of being robust to over-fitting and of not making strong assumption on the input features.

Table X displays the inference results for personality and hirability scores obtained using the different dimensionality reduction and regression techniques. For personality variables, only extraversion and openness to experience were inferred significantly more accurately than the baseline-average model, with R^2 values up to 0.27 for extraversion and 0.20 for openness to experience. Under the light of the analyses performed in Sections IV-E and VII, it is not surprising that these variables were the only ones accurately predicted: extraversion and openness to experience were the personality variables with the highest inter-rater agreements and with the largest number of correlated nonverbal cues. To contextualize these results, Biel *et al.* obtained inference results of $R^2 = .36$ for extraversion and $R^2 = .10$ for openness to experience in a dataset of video blogs [12]. This suggests that although video blogs seem to be a better-suited setting to predict extraversion, this task can also be achieved in video resumes, while more accurate results were achieved for openness to experience.

For hirability, all variables could be inferred significantly more accurately than the baseline-average model, with R^2 values up to 0.19 for overall impression, 0.15 for overall hirability, 0.12 for professional skills, 0.21 for social skills, and 0.20 for communication skills. Social and communication skills were the hirability variables with the most accurate prediction results, and were also the variables with the largest number of correlated cues (see Section VII). In comparison with the results obtained for the inference of hirability in employment interviews [35], R^2 values were lower, but the results were more significant due to the increase of data points by over an order of magnitude, showcasing the benefits of using large datasets.

In terms of regression methods, random forest with no dimensionality reduction consistently yielded the most accurate results. Dimensionality reduction methods did not increase the prediction accuracy, and this can be explained by the feature dimensionality of the original space ($D = 92$) compared to the number of data points.

Independently of the inference method and the analyzed social variable, the results obtained here show the feasibility of automatically inferring social impressions from video resumes to a moderate degree. Moreover, our initial hypothesis stating that nonverbal behavior can be used for the inference of personality and hirability variables holds. Despite the high variance in quality in the video resume dataset and the noise present in the features due to the error-prone nonverbal cue extraction step, positive inference results could be achieved. Furthermore, some of the results can be found in concordance with recent literature on nonverbal analysis of social videos [12].

B. Feature group analysis

In order to understand what cues were predictive of personality and hirability impressions beyond the correlation analysis performed in Section VII, we conducted a feature group analysis. Six feature groups were defined, based on the type of nonverbal behavior they were designed to represent, namely vocal, frontal face events, proximity, head motion,

TABLE X
PERFORMANCE (R^2 AND $RMSE$) FOR THE INFERENCE OF PERSONALITY AND HIRABILITY IMPRESSIONS USING DIFFERENT DIMENSIONALITY REDUCTION AND REGRESSION METHODS ($^*p < 10^{-3}$, $^\dagger p < 10^{-4}$ FOR $RMSE$, AND $p > 10^{-3}$ FOR VALUES WITH NO SYMBOLS). THE BEST ACHIEVED RESULT FOR EACH VARIABLE IS HIGHLIGHTED IN BOLD. $N = 882$.

	Extra.		**Agree.**		**Consc.**		**Neuro.**		**Open.**	
	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	0.00	0.61	0.00	0.39	0.00	0.45	0.00	0.39	0.00	0.51
All-Lasso	0.23	0.54†	0.01	0.39	0.00	0.45	-0.06	0.40	0.14	0.48†
All-Ridge	0.24	0.53†	0.03	0.38	0.02	0.45	0.00	0.39	0.17	0.47†
All-RF	**0.27**	**0.52**†	0.06	0.38	0.03	0.44	0.00	0.39	**0.20**	**0.46**†
Pval-Ridge	0.23	0.53†	0.04	0.38	0.00	0.45	-0.01	0.39	0.18	0.47†
Pval-Lasso	0.22	0.54†	0.04	0.38	-0.01	0.46	-0.03	0.40	0.17	0.47†
Pval-RF	0.25	0.53†	0.04	0.38	0.00	0.45	-0.06	0.40	0.18	0.46†
PCA-Ridge	0.23	0.53†	0.03	0.38	0.02	0.45	0.00	0.39	0.17	0.47†
PCA-Lasso	0.22	0.54†	-0.01	0.39	-0.00	0.45	-0.05	0.40	0.13	0.48*
PCA-RF	0.23	0.54†	0.04	0.38	0.02	0.45	0.01	0.39	0.17	0.47†

	Ov. Impression		**Ov. Hirability**		**Pro. Skills**		**Social Skills**		**Comm. Skills**	
	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	0.00	0.62	0.00	0.62	0.00	0.59	0.00	0.63	0.00	0.70
All-Lasso	0.12	0.58	0.09	0.59	0.08	0.56	0.11	0.59	0.09	0.66
All-Ridge	0.11	0.59*	0.09	0.60*	0.09	0.56*	0.12	0.59†	0.14	0.65†
All-RF	**0.18**	**0.56**†	**0.15**	**0.57**†	**0.12**	**0.55**†	**0.21**	**0.56**†	**0.20**	**0.62**†
Pval-Ridge	0.13	0.58†	0.09	0.60*	0.09	0.56*	0.12	0.59†	0.13	0.65†
Pval-Lasso	0.14	0.57†	0.11	0.59*	0.10	0.56	0.09	0.60	0.08	0.67
Pval-RF	0.17	0.57†	0.14	0.58†	0.11	0.56†	0.20	0.56†	0.18	0.63†
PCA-Ridge	0.12	0.58†	0.09	0.60*	0.09	0.56*	0.12	0.59†	0.14	0.65†
PCA-Lasso	0.10	0.59	0.08	0.60	0.05	0.57	0.08	0.60	0.08	0.67
PCA-RF	0.14	0.57†	0.10	0.59†	0.10	0.56†	0.13	0.59†	0.15	0.64†

overall motion (WMEI), and video statistics. To assess their predictive validity, each feature group was used individually in a regression task. Because random forest with no dimensionality reduction was consistently observed to be the most accurate method for the inference of personality and hirability impressions (see Section VIII-A), this regression method was used here.

Table XI displays the performance[1] (R^2 and $RMSE$) for the inference of personality and hirability scores using different feature groups and random forest with no dimensionality reduction, with D denoting the number of cues in each feature group. One can observe that despite the low number of vocal cues correlated with personality and hirability variables (see Section VII), the vocal feature group showed the highest predictive validity for all hirability variables and for extraversion and openness to experience. In particular, vocal cues alone achieved to explain 17% of the variance of extraversion and were marginal for the other variables. Similar results were previously obtained on video blogs, with vocal cues explaining 31% of the variance of the extraversion trait [12]. The relatively large ($D = 33$) number of cues contained in the vocal feature group fails to fully explain these results: head motion ($D = 24$) had a larger number of cues correlated with most variables, but showed poor predictive validity. To understand these results, we performed principal component analysis (PCA) on the feature groups and counted the number of components needed to explain 90% of the variance. For head motion, only 5 components achieved to explain over 90% of the variance, while 11 were necessary for vocal cues, showcasing the greater variety of the vocal feature group. We believe that this aspect at least partially explains the high predictive validity of this feature group. In addition to vocal cues, overall motion measured using weighted motion energy images, as well as basic video statistics, showed marginally good predictive validity (except for extraversion where it was more significant).

Overall, for personality and hirability variables, using all nonverbal cues yielded the best inference results. In other words, combining feature groups strongly improved the prediction accuracy. One hypothesis to interpret this observation is that each feature group explained a different part of the variance in the data, which added up when put in combination.

[1]Note that due to the random nature of random forest, the inference results may slightly differ between Tables X and XI.

IX. DISCUSSION

In this section, we contextualize our findings in relation with related work. We first discuss the similarities and differences between online conversational video resumes and face-to-face job interviews in Section IX-A. Then we compare our results to the ones obtained on video blogs in Section IX-B. Last, we discuss limitations and possible future work in Section IX-C.

A. Video resumes vs. job interviews

From the standpoint of job-seekers, online conversational video resumes and face-to-face employment interviews share one main common objective: to make a good first impression on potential employers in order to ultimately get hired for a position. In this context, the social constructs that matter most in both settings are the ones related to hirability impressions, as they are the ones determining the outcome. Despite these obvious similarities, online conversational video resumes differ from face-to-face job interviews on three levels.

The first main difference is that video resumes cannot be called *interactions* as such, as there is no conversational partner. The lack of conversational partner affected the results from this study compared to job interviews. In our previous

TABLE XI
PERFORMANCE (R^2 AND $RMSE$) FOR THE INFERENCE OF PERSONALITY AND HIRABILITY SCORES USING DIFFERENT FEATURE GROUPS AND RANDOM FOREST WITH NO DIMENSIONALITY REDUCTION ($^*p < 10^{-3}$, $^\dagger p < 10^{-4}$ FOR $RMSE$, AND $p > 10^{-3}$ FOR VALUES WITH NO SYMBOLS). THE BEST ACHIEVED RESULT FOR EACH VARIABLE IS HIGHLIGHTED IN BOLD. $N = 882$.

		Extra.		**Agree.**		**Consc.**		**Neuro.**		**Open.**	
	D	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	-	0.00	0.61	0.00	0.39	0.00	0.45	0.00	0.39	0.00	0.51
Vocal	33	**0.17**	**0.56**†	0.02	0.39	0.03	0.44	-0.02	0.40	**0.10**	**0.49**†
Frontal face events	9	0.02	0.60	-0.10	0.41	-0.17	0.49	-0.19	0.43	-0.01	0.52
Proximity	8	0.11	0.58*	-0.09	0.41	-0.08	0.47	-0.10	0.41	0.06	0.50
Head Motion	24	0.08	0.59	-0.05	0.40	-0.07	0.47	-0.06	0.40	0.08	0.49
Ov. motion	10	0.14	0.56†	-0.02	0.40	-0.02	0.46	-0.04	0.40	0.08	0.49
Video Stats	8	0.11	0.58†	-0.04	0.40	-0.04	0.46	-0.04	0.40	0.08	0.49
All-NVB	92	**0.28**	**0.52**†	0.05	0.38	0.03	0.44	0.00	0.39	**0.20**	**0.46**†

		Ov. Impression		**Ov. Hirability**		**Pro. Skills**		**Social Skills**		**Comm. Skills**	
	D	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$
Baseline-Avg	-	0.00	0.62	0.00	0.62	0.00	0.59	0.00	0.63	0.00	0.70
Vocal	33	**0.10**	**0.59**†	**0.07**	**0.60**	**0.06**	**0.57**	**0.09**	**0.60***	**0.13**	**0.65**†
Frontal face events	9	-0.08	0.65	-0.08	0.65	-0.06	0.61	-0.03	0.64	-0.06	0.72
Proximity	8	-0.00	0.62	-0.03	0.63	-0.07	0.61	0.01	0.62	-0.02	0.70
Head Motion	24	0.03	0.61	0.02	0.62	-0.01	0.59	0.06	0.61	0.04	0.68
Ov. motion	10	0.04	0.61	0.03	0.61	0.03	0.58	**0.09**	**0.60**	0.04	0.68
Video stats	8	0.07	0.60	0.06	0.60	0.02	0.58	0.08	0.60	0.07	0.67
All-NVB	92	**0.17**	**0.56**†	**0.15**	**0.57**†	**0.11**	**0.56**†	**0.20**	**0.56**†	**0.20**	**0.62**†

study on job interviews [35], we found that nonverbal cues based on speaking turns were predictive of hirability ratings. In this work, although cues related to speaking activity were extracted from video resumes, they were found to be of relatively low predictive validity, and we believe that this can be at least partly explained by the absence of a conversational partner. Additionally, interviewer visual cues were found to be predictive of hirability in [35]; these cues obviously could not be used in video resumes due to the lack of interaction partner.

The second major difference with job interviews is data quality. Job interviews datasets [35] [34] were recorded in laboratory settings where the environment is optimal (high-definition cameras with proper illumination, high-end microphones in no background noise), while online video resumes can potentially include music, improper illumination, low-quality audio recording, high compression rates, low video resolution, non-fixed cameras, improper framing, text displays and overlays, or highly textured background, multiple people on the video, unexpected user behavior. We believe that the decrease in overall data quality directly affected the feature extraction process, which could ultimately also have affected the inference results.

The third main difference with job interviews is that online video resumes are not limited with respect to the types of jobs subjects are applying to. This differs significantly from the related works where all subjects included in the datasets were applying for the same job, *e.g.* a marketing job [35]. We hypothesize that this can make a difference in terms of the relationships between the behavioral cues produced by the job applicant/online job-seeker and the hirability scores. For instance, the expected behavior for a person applying for a sales position will differ from someone looking for an engineering position. The differences across job types have however been addressed in psychology through meta-analytic studies [26], but to our knowledge, they have not been investigated in computing. One possible hypothesis is that the relationship between nonverbal cues and social constructs is conditioned by the type of jobs the person is looking for; if this was the case, aggregating multiple job types could have leveled out some relationships between nonverbal cues and social variables. This hypothesis deserves to be investigated in detail as future work, and the crowdsourced annotations of job types can be used to this end.

B. Video resumes vs. video blogs

In comparison with conversational video blogs [12], video resumes are similar in terms of setting: they are user-generated, include one person delivering a spoken message in front of a camera, and are posted on online video platforms such as YouTube. In this sense, video resumes also face the same challenges in terms of audio and video processing, and they also include some behavioral information emitted by the speaker. However, there are major differences between video blogs and video resumes, mainly in terms of objectives, content, and behavior of the communicators. In video resumes, online job seekers have the clear objective of making a good first impression on potential employers in order to be invited for a face-to-face job interview. In contrast, video blogs have a broader scope as they can be used for life documentary, daily interaction, e-learning, entertainment, or marketing [11]. We believe that these differences can affect the types of social constructs that come to the mind of external viewers when watching the videos.

One of the objectives of this work was to assess the level of agreement on several social constructs on this type of media, and our results show that obtaining a good inter-rater agreement is less obvious than in video blogs for all personality dimensions except for extraversion. It is well known in psychology that the specific situation has an impact on what traits and behaviors can be elicited and/or visible by others [31]. Given this fact, one hypothesis to explain the low ICC values can be the fact that video resumes do not elicit the expression of several personality traits.

C. Limitations and future work

One of the main limitations of this work is the simplicity of the visual features used for analysis. We used a frontal

face detector as a proxy for looking at the camera, but this constitutes a rather noisy estimation of the visual focus of attention as this feature can be dependent on how the face is positioned. One way to address this issue is to use head pose or gaze estimation methods, but the use of such techniques on this type of videos is a research problem that goes beyond the scope of this paper. The potentially low overall quality of some video resumes (in terms of framing, resolution, and lighting) constitutes the main challenge of analyzing gaze or head pose using automated methods.

Emotions, largely communicated through displays of facial expressions [19], has been shown to play a role in the formation of first impressions both in job interviews [33] and video blogs [13]. One possible avenue for future work would be to analyze the predictive validity of facial expressions in video resumes. The low quality of some video resumes constitutes an important challenge to reliably extract facial features, but this issue can be addressed by separating high and low quality video resumes, as described in the previous paragraph. Another potential issue is the fact that people are mostly speaking in video resumes, which can add complexity to the extraction of facial features, such as smiles.

Another limitation of this work resides in the low inter-rater agreement on some social dimensions obtained from crowdsourced annotations, such as the personality traits of agreeableness, conscientiousness, and neuroticism. The low ICC values limit the validity of the subsequent analyses. One possible option to address this issue would be to ask human resources professionals to re-annotate a subset of the video resumes.

X. Conclusion

In this work, we analyzed the formation of personality and hirability impressions on conversational video resumes. To the best of our knowledge, this work constitutes the first computational study on video resumes; it is also the first investigating video resumes from a nonverbal standpoint.

As a first step, we collected a dataset of 939 conversational English-speaking video resumes hosted on YouTube. Crowdsourced annotations of basic facts, demographics, perceived skills, and first impressions of hirability and personality were collected using Amazon Mechanical Turk, and most variables were found to be reliable upon analysis of the inter-rater agreement, not only suggesting that raters completed the HITs conscientiously, but also that the annotations of first impressions of job-related skills, hirability, and the personality variables of extraversion and openness to experience by unacquainted naïve judges was a feasible task. This in itself is a positive result regarding the use of crowdsourcing to obtain impressions at large scale. The personality variables of agreeableness, conscientiousness, and neuroticism, however, had low inter-rater agreement, suggesting that these traits were more difficult to rate in this setting. The analysis of the demographics showed that the job-seekers composing the video resume dataset were mainly young and applying for internship or junior positions. Demographics varied across job categories: the professional category (engineers, computer scientists, and scientists) was mainly composed of young Indian males, whereas hospitality was dominated by Caucasians, with an equal number of men and women. This is probably biased by the geographic origin of the posts, which YouTube does not give access to.

To understand the structure underlying the perceived skills, we conducted a clustering analysis, showing that skills could be grouped into three high-level clusters, namely professional skills, communication skills, and social skills. Nonverbal cues were automatically extracted from the visual and audio modalities to obtain a feature representation of the video resumes. As a first step, we performed a correlation analysis between nonverbal cues and the constructs of personality and hirability, and found that head motion, looking turns, proximity, and overall motion were correlated with extraversion, openness to experience, social skills, and communication skills, whereas only proximity and head motion cues were associated to overall hirability and overall impression. As a second step, we evaluated several regression methods for the inference of personality and hirability. Results demonstrated the feasibility of inferring hirability variables as well as extraversion and openness to experience in video resumes, achieving R^2 results up to 27% in a proper cross-validation experimental setting.

Several possible research directions can be considered for future work. First, the accuracy of the nonverbal cue extraction process could be improved despite the challenging nature of the video resumes. Second, more behavioral cues could be extracted, such as facial expressions, true gaze, head nods, or verbal content; moreover, non-behavioral cues could also be interesting to extract and analyze, such as the presence of music or changes of shots. Third, other inference tasks could be performed, such as the classification of job categories, or the inference of hirability and personality conditioned on the job categories. Last, all analyses in this chapter were conducted based on crowdsourced annotations of social variables by naïve raters. We strongly believe that a comparison with expert raters (*e.g.*, human resources professionals) could be beneficial for a deeper understanding of how first impressions are made in organizational settings; such gold-standard annotations would allow us to assess and compare the predictive validity of automatically nonverbal cues and crowdsourced annotations.

Acknowledgments

This work was funded by the UBIMPRESSED project of the Sinergia interdisciplinary program of the Swiss National Science Foundation (SNSF). We would like to thank (1) Dr. Biel for the help he gave to the design and analysis of the experiment; and (2) the crowdworkers for contributing their impressions.

References

[1] Amazon Mechanical Turk. Available: https://www.mturk.com [online].
[2] American Time Use Survey. Available: http://www.bls.gov/tus/tables/a4_1112.pdf [online].
[3] FFmpeg: A complete, cross-platform solution to record, convert and stream audio and video. Available: https://www.ffmpeg.org/ [online].
[4] MIT speech feature extraction code. Available: http://groupmedia.media.mit.edu/data.php [online].

[5] Youtube data API (v3). Available: https://developers.google.com/youtube/v3/ [online].

[6] 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.

[7] S. Basu. *Conversational Scene Analysis*. PhD thesis, Massachussetts Institute of Technology, 2002.

[8] L.M. Batrinca, B. Lepri, N. Mana, and F. Pianesi. Multimodal Recognition of Personality Traits in Human-Computer Collaborative Tasks. In *Proc. Int. Conf. on Multimodal Interaction*, 2012.

[9] L.M. Batrinca, N. Mana, B. Lepri, F. Pianesi, and N. Sebe. Please, Tell Me About Yourself: Automatic Personality Assessment using Short Self-Presentations. In *Proc. Int. Conf. on Multimodal Interaction*, 2011.

[10] 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. Int. Conf. on Web and Social Media*, 2011.

[11] J.-I. Biel and D. Gatica-Perez. VlogSense: Conversational Behavior and Social Attention in YouTube. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2(3):1–20, 2010.

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

[13] J.-I. Biel, L. Teijeiro-Mosquera, and D. Gatica-Perez. FaceTube: Predicting Personality from Facial Expressions of Emotion in Online Conversational Video. In *Proc. Int. Conf. on Multimodal Interaction*, 2012.

[14] J.-I. Biel, V. Tsiminaki, J. Dines, and D. Gatica-Perez. Hi YouTube! Personality impressions and verbal content in social video. In *Proc. Int. Conf. on Multimodal Interaction*, 2013.

[15] L. Breiman. Random Forests. *Machine Learning*, pages 1–35, 2001.

[16] 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):802, 2007.

[17] T. DeGroot and J. Gooty. Can Nonverbal Cues Be Used to Make Meaningful Personality Attributions in Employment Interviews? *Business and Psychology*, 24(2):179–192, 2009.

[18] A. Doyle. Skills List for Resumes. Available: http://jobsearch.about.com/od/list/fl/list-of-skills-resume.htm [online].

[19] P. Ekman. *Emotion in the Human Face*. Cambridge University Press, 2nd edition, 1982.

[20] J.L. Fleiss, B. Levin, and M.C. Paik. The Measurement of Interrater Agreement. In *Statistical Methods for Rates and Proportions*, pages 598–626. John Wiley and Sons, 3rd edition, 2003.

[21] R.J. Forbes and P.R. Jackson. Non-Verbal Behaviour and the Outcome of Selection Interviews. *Occupational Psychology*, 53(1):65–72, 1980.

[22] D. Gatica-Perez. Automatic Nonverbal Analysis of Social Interaction in Small Groups: A Review. *Image and Vision Computing*, 27(12):1775–1787, 2009.

[23] S. Gosling. A Very Brief Measure of the Big-Five Personality Domains. *Research in Personality*, 37(6):504–528, 2003.

[24] S.D. Gosling, S. Gaddis, and S. Vazire. Personality Impressions Based on Facebook Profiles. In *Proc. Int. Conf. on Web and Social Media*, 2007.

[25] A.M.F. Hiemstra. *Fairness in Paper and Video Resume Screening*. PhD thesis, Erasmus University Rotterdam, Netherlands, 2013.

[26] A.I. Huffcut, J.M. Conway, P.L. Roth, and N.J. Stone. Identification and Meta-Analytic Assessment of Psychological Constructs Measured in Employment Interviews. *Journal of Applied Psychology*, 86(5):897–913, 2001.

[27] A.S. Imada and M.D. Hakel. Influence of Nonverbal Communication and Rater Proximity on Impressions and Decisions in Simulated Employment Interviews. *Applied Psychology*, 62(3):295–300, 1977.

[28] 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.

[29] J.F. Kelly and E.H. O'Brien. Using Video Resumes to Teach Deaf College Students Job Search Skills and Improve Their Communication. *American Annals of the Deaf*, 137(5):404–410, 1992.

[30] K.J. Kemp, L.M. Bobbitt, M.B. Beauchamp, and E.A. Peyton. Using One-Minute Video Résumés as a Screening Tool for Sales Applicants. *Marketing Development and Competitiveness*, 7(1):84–92, 2013.

[31] M.L. Knapp and J.A. Hall. *Nonverbal Communication in Human Interaction*. Wadsworth, Cengage Learning, 7th edition, 2009.

[32] B. Lepri, R. Subramanian, K. Kalimeri, J. Staiano, F. Pianesi, and N. Sebe. Connecting Meeting Behavior with Extraversion - A Systematic Study. *IEEE Transactions on Affective Computing*, 3(4):443–455, 2012.

[33] S.P. Levine and R.S. Feldman. Women and Men's Nonverbal Behavior and Self-Monitoring in a Job Interview Setting. *Applied Human Resource Management Research*, 7(1):1–14, 2002.

[34] I. Naim, M.I. Tanveer, D. Gildea, and M.E. Hoque. Automated Prediction and Analysis of Job Interview Performance: The Role of What You Say and How You Say It. In *Proc. Int. Conf. on Automatic Face and Gesture Recognition*, 2015.

[35] L.S. Nguyen, D. Frauendorfer, M. Schmid Mast, and D. Gatica-Perez. Hire Me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Transactions on Multimedia*, 16(4):1018–1031, 2014.

[36] R.E. Nisbett and T. Decamp Wilson. The Halo Effect: Evidence for Unconscious Alteration of Judgments. *Personality and Social Psychology*, 35(4):250–256, 1977.

[37] F. Pianesi, N. Mana, A. Cappelletti, B. Lepri, and M. Zancanaro. Multimodal Recognition of Personality Traits in Social Interactions. In *Proc. Int. Conf. on Multimodal Interfaces*, page 53, 2008.

[38] C. Piotrowski and T. Armstrong. Current Recruitment and Selection Practices: A National Survey of Fortune 1000 Firms. *North American Journal of Psychology*, 8(3):489–496, 2006.

[39] J.A. Rolls and M. Strenkowski. Video Technology: Resumes of the Future. In *World Conference on Cooperative Education*, 1993.

[40] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez. A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups. *IEEE Transactions on Multimedia*, 14(3):816–832, 2012.

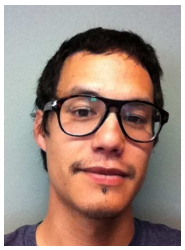
[41] D. Sanchez-Cortes, J.-I. Biel, S. Kumano, J. Yamato, K. Otsuka, and D. Gatica-Perez. Inferring Mood in Ubiquitous Conversational Video. In *Proc. Int. Conf. on Mobile and Ubiquitous Multimedia*, 2013.

[42] P.E. Shrout and J.L. Fleiss. Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

[43] C. Smith. By the Numbers: 120 Amazing LinkedIn Statistics. Available: http://expandedramblings.com/index.php/by-the-numbers-a-few-important-linkedin-stats/ [online].

[44] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Royal Statistical Society*, 58(1):267–288, 1996.

[45] P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*, 2001.



Laurent Son Nguyen is a postdoctoral researcher at the Social Computing Group at Idiap Research Institute, Switzerland. He obtained his Ph.D. in 2015 from École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland on the automated analysis of human behavior in job interviews. He is mainly interested in using computational approaches to understand the formation of first impressions in face-to-face interactions, online videos, and social media images.

Daniel Gatica-Perez is Head of the Social Computing Group at Idiap and Professeur Titulaire at the École Polytechnique Fédérale de Lausanne (EPFL) in Switzerland. His research interests include social computing, social media, ubiquitous computing, and crowdsourcing. He has served as an associate editor of the IEEE Transactions on Multimedia. He is a member of the IEEE.