Facing Employers and Customers: What Do Gaze and Expressions Tell About Soft Skills?

Skanda Muralidhar
Idiap, EPFL
Switzerland
smuralidhar@idiap.ch

Rémy Siegfried
Idiap, EPFL
Switzerland
remy.siegfried@idiap.ch

Jean-Marc Odobez
Idiap, EPFL
Switzerland
odobez@idiap.ch

Daniel Gatica-Perez
Idiap, EPFL
Switzerland
gatica@idiap.ch

ABSTRACT
Eye gaze and facial expressions are central to face-to-face social interactions. These behavioral cues and their connections to first impressions have been widely studied in psychology and computing literature, but limited to a single situation. Utilizing ubiquitous multimodal sensors coupled with advances in computer vision and machine learning, we investigate the connections between these behavioral cues and perceived soft skills in two diverse workplace situations (job interviews and reception desk). Pearson’s correlation analysis shows a moderate connection between certain facial expressions, eye gaze cues and perceived soft skills in job interviews ($r \in [-30, 30]$) and desk ($r \in [20, 36]$) situations. Results of our computational framework to infer perceived soft skills indicates a low predictive power of eye gaze, facial expressions, and their combination in both interviews ($R^2 \in [0.02, 0.21]$) and desk ($R^2 \in [0.05, 0.15]$) situations. Our work has important implications for employee training and behavioral feedback systems.

ACM Classification Keywords
I.4.m. Image Processing and Computer Vision: Miscellaneous; J.4 Social and Behavioral Sciences

Author Keywords
First impressions; hirability; job performance; eye gaze; facial expressions; social computing; hospitality; multimodal interaction;

INTRODUCTION
Eye gaze and facial expressions are the foundations of communication and social interaction in humans. They are also the easiest nonverbal cues to signal interest; a person who avoids eye contact doesn’t want to be bothered while if he/she smiles and makes eye contact, it indicates his/her willingness to interact. Previous works have explored the interconnections between eye gaze, facial expressions, and various social variables including from hirability [15, 21], sales performance [26], dominance [20] and leadership [34] in the context of a single situation in workplace. In this work, using ubiquitous multimodal sensors and advances in machine learning, we investigate the connections between these nonverbal behavioral cues and first impressions across two diverse workplace settings: job interviews and reception desk.

Literature in psychology has shown that eye gaze and facial expressions play a role in the formation of first impressions in workplaces [24]. In the context of employment interviews, Forbes et al. [15] reported that successful applicants in job interviews made more direct eye contact, produced more facial expressions, smiled and nodded more than unsuccessful ones. In the context of customer-centric service too, these behavioral cues have been reported to be important. In particular, Tsai et al. [45] found that employees’ smiling and eye contact were related to the customers’ greater willingness to revisit the store. Other studies have shown links between frequent and longer eye contact and increased likability [5], satisfaction [23], credibility and trust [18]. Psychology literature traditionally relies on manual coding of eye gaze and facial expressions, which is an expensive, time consuming and labor-intensive process. With advances in mobile and ubiquitous sensors, capturing eye gaze and facial expressions are becoming inexpensive and common (e.g. Apple’s Animoji). Using such advances, eye gaze and facial expressions have been investigated in computing literature. In the context of workplace settings, various social variables including dominance [22, 19], leadership [35], personality [3] have been investigated from the perspective of these cues. In the context of job interviews, Chen et al. [9] investigated the eye gaze, facial expressions and head postures. Using a dataset of 36 videos, the authors reported a correlation of $r = 0.36$ between expert hirability ratings and visual features represented as “visual words”. 7 classes of facial expressions...
and their connections to interview performance was explored by Chen et al. [10] using 20 structures interviews. The authors reported a positive correlation ($r = 0.52$) between joy and hirability, while anger was negatively correlated ($r = -0.44$).

Due to the challenges of collecting and analyzing multimodal data, most existing automated prediction frameworks focus on a single setting. In this work, we study the connections between eye gaze, facial expressions and first impressions in the context of two workplace settings; job interviews and hotel reception desk. Specifically, we investigate the connections between eye gaze, facial expressions and perceived hirability in interviews and perceived performance on the job. Towards this, we utilize a previously collected data corpus consisting of 169 videos each from job interviews and desk interactions.

The contribution of this work are: (1) We extract eye gaze and facial expressions using state-of-the-art computer vision and machine learning. (2) In the job interview setting, a Pearson’s correlation analysis revealed (a) moderate positive correlation between eye gaze and perceived hirability ($r = 0.28$), (b) negative correlation between neutral face and perceived hirability ($r = -0.28$). (3) In the reception desk setting, a Pearson’s correlation analysis revealed moderate positive correlation between (a) eye gaze and perceived performance ($r = 0.36$), (b) anger and perceived performance ($r = 0.24$). (4) Our computational framework to automatically infer perceived soft skills showed low to moderate inference performance for interviews ($R^2 \in [0.02, 0.21]$) and desk ($R^2 \in [0.05, 0.15]$) situations.

Our results demonstrate the feasibility of quantifying such behavioral cues and their contributions to perceived soft skills in various settings, which have previously been reported mainly qualitatively in psychology literature using advances in ubiquitous multimodal sensors and machine learning.

**DATASET**

**Data Collection**

This dataset consists of 169 interactions previously collected by us [27], role-played by students of a hospitality school in two workplace settings; job interview and reception desk (total of 338 videos). Both the settings were recorded synchronously using (1) two Kinect cameras (one for each protagonist) at 30 frames per second; (2) a microphone array placed at the center of the table recording audio at 48kHZ. Due to challenges in feature extraction due to non-frontal body pose, this work uses 161 videos for interview and 153 for desk. This dataset was annotated by us previously [27] for various social variables (Table 1). The annotations for interview videos were done by 5 raters, while for desk there were 3 raters. Both groups were mutually exclusive and consisted of students, who were paid approximately 20 USD per hour for this work. The raters watched the first two minutes of the videos and annotated the variables on a seven-point Likert scale. Use of two-minute segments, also known as thin slices, follows existing literature [2, 33]. Details of the data collection can be found in [27].

Table 1 presents the inter-rater agreement and their descriptive statistics. Agreement between raters was measured using Intraclass Correlation Coefficient (ICC) [37], a common metric used in psychology. Since annotations were done using a sample of annotators and each annotator judged all videos, we use ICC($2,k$). The tables show that the inter-rater agreement was acceptable with ICC($2,k$) > 0.5. Specifically, we observe moderate to high agreement between the raters with ICC($2,k$) $\in [0.52, 0.77]$ for interview videos and ICC($2,k$) $\in [0.60, 0.77]$ for reception desk. Further, we observe that the mean values of the annotated variables are centered on the positive side of the Likert scales (Mean $\geq 4$), suggesting that the participants were generally perceived positively by both groups of annotators.

**FEATURES EXTRACTED**

In the section we outline the extraction of eye gaze and facial expression cues. A number of statistics, including count, mean, median, standard deviation, minimum, and maximum were computed from these cues and used as features.

**Eye Gaze**

We extract eye gaze using an off-the-shelf method to compute participants’ gaze (i.e. line of sight in 3D space) from a Kinect v2 sensor [16]. Taking advantage of the context (i.e. a human-human interaction), we improved its outputs using a subject adaptation method that computes the gaze estimation bias based on speaking turns [38]. An overview of this method is presented in Figure 2 and follows the intuition that people are more likely to look at the person who is speaking during a conversation. The gaze errors corresponding to the frames where the other person is speaking and where the estimated gaze and head pose make a glance at the speaker possible are stored. Then, the bias is estimated using the Least Median Square (LMedS) estimator, so that remaining outliers frames (i.e. frames where the subject is not looking at the other person despite the fact that he/she is talking) are removed. Finally, the bias can be subtracted from the gaze signal, which allows us to compute the Visual Focus of Attention (VFOA) with an

---

**Table 1: List of perceived variables manually annotated for both situations, along with their ICC($2,k$) and means.**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Job Interview</th>
<th>Reception Desk</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Professional Skills</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Competent (compe)</td>
<td>0.56</td>
<td>0.69</td>
</tr>
<tr>
<td>Motivated (motiv)</td>
<td>0.52</td>
<td>0.63</td>
</tr>
<tr>
<td><strong>Social Skills</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Positive (posit)</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>Sociable (socia)</td>
<td>0.57</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>Communication Skills</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clear (clear)</td>
<td>0.67</td>
<td>0.66</td>
</tr>
<tr>
<td>Persuasive (persu)</td>
<td>0.69</td>
<td>0.72</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Performance (peImp)</td>
<td>–</td>
<td>0.77</td>
</tr>
<tr>
<td>Hirability (hire)</td>
<td>0.69</td>
<td>4.11</td>
</tr>
</tbody>
</table>

---

**Figure 1:** Snapshot of the situations studied in this work: (a) job interview, (b) reception desk.
higher accuracy without the need to for manual calibration or annotation. This method allowed to reach a mean error of 7.64° on annotated subsets of the interview and desk datasets.

Using this data, we extract the various eye gaze cues, defined as number and duration of events when participant was gazing at the protagonist, from both client and participant. These include gaze while speaking (GWS; number and duration of looking at the protagonist and speaking), gazing while listening (GWL; number and duration of looking at the protagonist and not speaking), visual dominance ratio (VDR defined as ratio of GWS and GWL [14]), and mutual gaze.

Facial Expressions

Literature in psychology indicates connections between facial expressions and impressions of dominance [30, 43], warmth [4] and emotional distance [7]. We extract facial expressions displayed from using “off-the-shelf” Emotion API [12] of Microsoft Azure cognitive services [42, 41]. Such cloud-based services have been used in literature to study diverse social, and political issues [36] including cyber-bullying [39] and public health images [17].

Microsoft Azure Emotion API recognizes 8 facial expressions of emotions (happiness, sadness, surprise, anger, fear, contempt, disgust and neutral). As first step, we identify segments in the videos where the protagonist is not speaking so as to avoid mis-classifications of emotion classes (due to movement of facial muscles and lips). From these segments, we extracted frames at 5 fps, which are then input sequentially to the Emotion API to maintain temporal continuity. If the API detected a face in the frame, its returns confidence values across the 8 emotion types normalized to 1. Else if a face was not found, the API returns 0 for all values. Frames with no faces were filtered before computing various statistics was computed from this 8-dimensional vector and used as features.

Other Features

To further understand the impact of gaze and facial expressions, we combine them with other visual and auditory cues previously extracted by us in [27, 28]. Visual features consists of (1) Overall visual motion captures the total amount of visual movement displayed [6]; and (2) Head nods were extracted using a 3D face centered method [11]. Apart from simple head nods, nodding while speaking was extracted by synchronizing speaking activity with head nod. Auditory features include (1) Speaking activity includes speaking time, speaking turns, pauses, short utterances, and silence; and (2) Prosody includes pitch, speaking rate, spectral entropy, energy, voicing rate, and time derivative of energy, extracted using [8, 32].

Table 2: Range of Pearsons correlation between eye gaze, facial expressions and social variables in the interview (N = 161) and reception desk (N = 153) dataset. *p < 0.001; **p < 0.01; ***p < 0.05.

CORRELATION ANALYSIS

In the first step, we conduct a Pearson’s correlation analysis between perceived soft skills scores, gaze and facial expressions for the two settings (Table 2). In the interview setting, we observe low to moderate correlations between eye gaze, facial expressions and perceived soft skills. Specifically, we observe that applicants who had greater VDR (i.e. amount of visual dominance) and presented a happy expression were perceived to be more hirable than applicants who presented a neutral expression and displayed lower VDR (i.e. looked less at the interviewer while speaking). Our results concur with literature, which report positive correlation between eye gaze and hirability [15, 21, 1], and happy and hirability [21, 10].

Similarly, in the desk setting moderate correlations between eye gaze, facial expressions and perceived soft skills were observed. In particular, we observe that participants who held client’s gaze for longer duration while speaking, displayed greater VDR and nonverbal immediacy (by mirroring anger the clients displayed), were perceived to be better performing than participants who did not display these cues. These connections are supported in literature. It has been shown that even in situations when the client is dissatisfied, greater eye contact leads to enhanced perception of credibility [44]. Soderlund et al. [40] reported that dissatisfied client’s assessment of employee’s emotional state affects their own emotional state, which, in turn, impacts customer’s level of satisfaction.

Interestingly, the protagonists’ (interviewer and client respectively) gaze and expression cues were not correlated to perceived hirability and performance. This is in contrast to other nonverbal cues (like speaking time) reported in literature for interviews [29, 27] and reception desk [28].

INFERENCE ANALYSIS

The inference of perceived variables was defined as a regression task and was evaluated using random forest (RF) algorithm from the CARET package [25] in R. Hyper-parameters (i.e., number of trees) were automatically tuned by using an inner 10-fold cross-validation (CV) on the training set. The final machine inferred scores was obtained by leave-one-video-out
We see that gaze and facial expressions have a low-moderate effect on interview outcomes. Chen et al. [9] evaluated impact of eye gaze and nonverbal behavior on perceived performance of salespersons. They reported significant effect of eye gaze on perceived believability, tactfulness and empathy. DeGroot et al. [13] investigated perceived performance of 110 managers in a news-publishing company, reporting a correlation of $r = 0.14$ between visual cues displayed and performance.

Our experimental framework shows that eye gaze and facial expressions have an impact on perceived soft skills, especially in customer-centric situations. We believe that this could be due to job interviews being a more predictable setting, with the applicants seated and hence restrictive in terms of behavioral display. In comparison, reception desk is a more unpredictable and spontaneous interaction with greater need for nonverbal communication to handle an unpleasant client.

**CONCLUSION**

This study used the links between eye gaze, facial expressions and perceived soft skills in two diverse workplace settings. Towards this, we utilized a previous dataset of 338 interactions in job interview and reception desk settings. First, using advances in multimodal sensor, computer vision and machine learning, we automatically extracted eye gaze and facial expressions. A Pearson’s correlation analysis showed low-to-moderate correlations between eye gaze, facial expressions and perceived soft skills in interviews and moderate correlations in the desk setting. A computation framework to infer perceived soft skills (formulated as a regression task) using these features showed low performance for both the settings. Furthermore, fusion of these cues with other visual and auditory features showed high performance in interview performance for some variables. Specifically, for Positive ($R^2 = 0.21$) followed by Social ($R^2 = 0.18$), Hirability ($R^2 = 0.13$), and Persuasive ($R^2 = 0.11$).

As a last step, we fused all the visual features with auditory behavioral cues. We observe a moderate improvement in inference performance (as compared to the baseline) for all the variables except Social and Motivated. Specifically, we observe improved inference for Positive ($R^2 = 0.34$), Persuasive ($R^2 = 0.33$), Social and Performance (both $R^2 = 0.32$).

These results are in accordance with those reported in literature. Leigh et al. [26] evaluated impact of eye gaze and other nonverbal behavior on perceived performance of salespersons. They reported significant effect of eye gaze on perceived believability, tactfulness and empathy. DeGroot et al. [13] investigating perceived performance of 110 managers in a news-publishing company, reported a correlation of $r = 0.14$ between visual cues displayed and performance.

**ACKNOWLEDGMENTS**

This work was partially funded by the Swiss National Science Foundation (SNSF) through the UBImpressed project.
REFERENCES


