Spotting Prejudice with Nonverbal Behaviours

Andrea Palazzi*, Simone Calderara
Dept. of Engineering, Univ. of Modena and Reggio Emilia

Nicola Bicocchi
Dept. of Engineering, Univ. of Modena and Reggio Emilia

Loris Vezzali, Gian Antonio di Bernardo
Dept. of Education, Univ. of Modena and Reggio Emilia

Franco Zambonelli
Dept. of Sciences and Methods for Engineering, Univ. of Modena and Reggio Emilia

Rita Cucchiara
Dept. of Engineering, Univ. of Modena and Reggio Emilia

ABSTRACT

Despite prejudice cannot be directly observed, nonverbal behaviours provide profound hints on people inclinations. In this paper, we use recent sensing technologies and machine learning techniques to automatically infer the results of psychological questionnaires frequently used to assess implicit prejudice. In particular, we recorded 32 students discussing with both white and black collaborators. Then, we identified a set of features allowing automatic extraction and measured their degree of correlation with psychological scores. Results confirmed that automated analysis of nonverbal behaviour is actually possible thus paving the way for innovative clinical tools and eventually more secure societies.

CCS Concepts

•**Human-centered computing** → **Ubiquitous and mobile computing theory, concepts and paradigms;** *Empirical studies in collaborative and social computing; Ambient intelligence;* •**Computing methodologies** → *Machine learning;*

Keywords

Social Interactions; Prejudice; Nonverbal Behaviours

1. INTRODUCTION

Intergroup nonverbal behaviours (INVB) represent a relevant part of the human communication process. INVB include a range of nonverbal behaviours that individuals enact when interacting with members of a different group. Examples of nonverbal behaviours generally investigated by literature are body movements/gestures, interpersonal distance, eye gaze, nodding, speaking time [17].

*All authors can be contacted at name.surname@unimore.it

UbiComp '16, September 12-16, 2016, Heidelberg, Germany

DOI: http://dx.doi.org/10.1145/2971648.2971703

However, despite their practical relevance and theoretical value, INVB are an understudied topic so far. Furthermore, previous research on INVB rely on costly or invasive procedures, mainly involving subjective annotations of video-recorded interactions. Recent technological advancements such as wearable sensing devices and RGB-depth cameras provided the foundational basis for capturing objective measures and indices (e.g., interpersonal distance, gestures) in a fully automatic and continuous way. This could: sensibly reduce subjective influences introduced by human coders (people annotating psychological experiments), and increase the number of measures (derived from the video, audio, or physiological domains) eventually undertaken at the same time and their temporal resolution [11], [21], [20]. This is not only a change of methodology, but could represent the beginning of a revolution in both clinical and social psychology. Indeed, for the first time scientists can collect accurate and objective measures not subject to human biases. Furthermore, these measures can be computed in real-time thus allowing on time, continuous, behavioural feedbacks and interventions [2], [16], [28]. Finally, recent sensors can gather information on features that are completely inaccessible for human coders, such as participants' biometric parameters [7], [15], [12], [25].

The contribution of this work is threefold: *(i)* we show that recent technologies are capable of capturing and recognising prejudice from nonverbal behaviours using a relatively limited set of features, *(ii)* we evaluate these technologies and show that, under specific assumptions, automated annotation of psychological experiments can be faster, cheaper, and more accurate than human annotation. Furthermore, *(iii)* using our approach we discovered an innovative feature, never explicitly hypothesised in previous psychological research, that proved be to significant in INVB classification and that can be actually collected only using machines rather than human evaluators.

The rest of the paper is organised as follows. In Section 2 we discuss related work in the field. In Section 3 we describe our study both in terms of devices and methodology used. The behavioural features we decided to consider are detailed in Section 4. In Section 5 we discuss the results of the study showing which features can be used to infer prejudice from nonverbal behaviours and their degree of correlation with psychological indexes. Using these data, we also show how to automatically recognize interactions characterised by an high degree of prejudice. Finally, Section 6 concludes the

paper.

2. RELATED WORK

Intergroup nonverbal behaviours are receiving increasing attention: understanding their development, consequences, and internal constructs they continuously represent is a compelling task. So far, nonverbal behaviours have been typically examined by means of human coders. Among the most common indices to be observed there are eye gaze, interpersonal distance, body orientation, self-touch, gestures, smiling, speaking time. However, coders perceptions are subjective and often invasive or expensive to obtain.

Nevertheless, recent advances in sensing and monitoring techniques are starting to make the automated coding of human behaviour possible [6], [5]. Among all the new digital technologies, certainly the diffusion of Microsoft Kinect [29] had the biggest impact. Indeed, despite being relatively cheap, the Kinect sensor can provide a large set of fairly accurate data useful for detecting nonverbal behaviours, such as people's postures, gestures and facial expressions. Furthermore, its unobtrusiveness is an important feature that makes it suitable for many applications in this field. The success of automated approaches as a substitute for human coding is testified by the growing number of researches in this field. In the following we discuss the most relevant ones.

In 2012 Burba et al [8] used Microsoft Kinect to measure subtle nonverbal behaviour features of users interacting with virtual human agents in order to estimate their psychological states. In particular they calculated the respiratory rate, estimated by measuring the visual expansion and contraction of the user chest cavity and a specific type of fidgeting behaviour, known as leg jiggling, by measuring high-frequency vertical oscillations of the user's knees.

In 2013 Lee et al. [18] developed a computational model for recognising interpersonal trust in social interactions. Their research is built upon the hypothesis that nonverbal behaviour can be predictive of the level of trust. In particular some gestural cues like leaning-backward, face-touching, hand-touching, and crossing-arm are commonly associated to lower levels of trust, while positive gestural cues as leaning-forward, having arms-in-lap, and open-arms can predict higher levels of trust. The same year Microsoft Kinect has also been successfully employed to identify nonverbal predictors of depression and post-traumatic stress disorders [24].

In 2014 Won at al. [26] used Microsoft Kinect to record a set of teacher-learner interactions to predict the learning performance by analysing nonverbal behaviours (gestures) that took place during the lecture. In a later study [27], the same authors demonstrated a relevant correlation between nonverbal synchrony of two people collaborating in a creative task and their success in the same task.

Recently, Bharathi et. al. [4] investigated the impact of automatic social behaviour characterisation in a gamification context. However, there is still a lack of technological equipment allowing the measurement of non-discrete aspects of nonverbal behaviours (e.g., interpersonal distance, body inclination) over the course of an interaction. Such advancement would allow calculating objective indices and testing predictions that are difficult or impossible to examine with current procedures. Furthermore, it would provide real-time, continuous monitoring of nonverbal behaviours enabling scientists to investigate a new spectrum of applicative scenarios [17] .

To the best of our knwoledge, the study we present here is the first and most extensive regarding automated detection of INVB so far. In fact, we collected data using both visual and physiological sensors. Furthermore, we also validated our results against an actual social experiment based on the presented methodology.
The corpus of collected data, comprising both 3D skeletons and phycological annotations is publicly available for download here: http://imagelab.ing.unimore.it/spotting-prejudice.

3. STUDY DESIGN

We designed a study on prejudice towards black people in which 60 participants were requested to talk about the same topics with both white and black peers collaborating with us. All interactions have been recorded with several sensors to extract as many features as possible. To minimise experimental artefacts, we gave participants no constraint on how to behave during the interactions. For each participant, we tested both *implicit* (largely outside conscious awareness) and *explicit* prejudice (of which individuals are clearly aware).

From a psychological perspective, we focused on measuring the degree of prejudice represented by the implicit-association test (IAT) score [13]. It has been designed to detect the strength of a person's automatic association between mental representations of objects (concepts) in memory. More specifically, in the IAT participants had to categorize, as fast and accurate as possible, black faces, white faces, positive words and negative words by pressing W or P on the keyboard. In one block of 40 trials (Block A) black faces and positive words shared the same response key (e.g., W) while white faces and negative words were associated with the other key (e.g., P). In a second block (Block B, 40 trials), these associations were inverted, namely, one key (e.g., W) was assigned to black faces and negative words and the other key (e.g., P) was employed to categorize white faces and positive words. Implicit prejudice is measured by computing the difference between the two blocks ($BlockA - BlockB$) in a way that higher scores express stronger automatic prejudice. For the detailed scoring procedure, see [14].

3.1 Participant Population

We started with 60 participants with ages comprised between 20 and 25 ($\mu = 21.15, \sigma = 1.89$). The vast majority of them are studying engineering at our University and were recruited - without economic incentives - using a public online form. After the experiment, 28 participants were removed for one or more of the following reasons: *(i)* participant knew at least one of our collaborators, *(ii)* participant were themselves black, *(iii)* equipment failure or missing data [22], [19]. This left a set of 32 participants (20 males and 12 females). Before the beginning of the experiment all participants signed an informed-consent and gave their agreement to the treatment of their personal data.

3.2 Devices used

As mentioned above, to capture participant's behaviour in an unobtrusive way, we used a Microsoft Kinect V2. The Kinect was physically hanged on the top of the interaction platform. A GoPro camera was also used to record all the interactions from a different point of view. The GoPro camera was hanged on the other side of the stage with respect to the Kinect sensor. Considering the shorter distance between

Figure 1: Each participant (in red) met both white (left) and black (right) collaborators (in yellow). Collaborators were already on the scene when the participant arrived, so their mutual distance has been always set by the participant. Indeed, both people were free to move wherever they wanted within the scene.

this camera and the interlocutors, we extracted the audio signal from the GoPro camera. Furthermore, to collect participant's biometric data, we asked all participants to wear a Shimmer GSR device [9] during the interactions. Shimmer sensor is worn like a watch and is extremely light, so does not compromise the spontaneity of the interaction. From this device we acquired two biometric measures, namely PPG and GSR signals, related to heart rate and emotional arousal respectively.

3.3 Experimental procedure

We adapted a standard procedure used in psychological research to assess the predictive role of explicit and implicit prejudice on nonverbal behavior with some differences: (a) for practical reasons, explicit prejudice was assessed in the experimental session and not in a pre-test; (b) for the sake of examining features such as interpersonal distance and space volume (see below), participants were asked to stand up instead of sitting; (c) in order to better understand the relative effects of explicit and implicit prejudice depending on contextual conditions, participants were asked to discuss a race relevant topic in addition to a neutral topic [10]. In the following, the complete procedure is detailed.

Each participant, at arrival, has been taken in a first room. Here, each participant has been asked by researchers to fill two different questionnaires. The first questionnaire (Q1) contained questions about her perceptions of black people. This questionnaire aims at gathering information about the so called *explicit prejudice*. We say that this measure is explicit because while answering the questionnaire the participant is aware of the answer provided regarding attitudes towards black people. The second questionnaire, instead, consists in an implicit-association test (IAT) aiming at discovering the *implicit prejudice* of the participant towards black people. The peculiarity of this test lies in that the result is not computed from the answers themselves, but from the time the participant took to answer each question. In this way, it is almost impossible for the participant to disguise her opinion.

After filling the two questionnaires, the participant has been taken to a second room where the recording platform has been arranged. Here, the participant met the partner for the first interaction. Researchers informed both persons that the interaction was composed by two different conversations during three minutes each. More specifically, one regarded immigration (*salient* topic) while the other was frivolous (*nonsalient* topic). As soon as researchers left the room, the participant and the collaborator started their conversations. During the dialogue, they were completely free to move wherever they wanted within the recording stage.

After six minutes, both the participant and the collaborator have been accompanied in separate rooms to fill another questionnaire (Q2) about their impressions and feelings. This questionnaire was significantly different for the participant and the collaborator. Indeed, the participant had to answer questions about her own feelings during the interaction, while the collaborator had to guess which were the dominant emotions felt by the participant during their interaction. Once completed the second questionnaire, the participant has been taken back in the recording room, where another collaborator for the second interaction was waiting. The second collaborator was white if the first one was black and vice versa, so the participant always talked about the same two topics with people of different race. The second interaction was identical to the first one. After other two conversations of three minutes each, the second interaction ended and both the participant and the collaborator were accompanied in separate rooms to fill their questionnaires.

3.4 Psychological measures

After each participant completed the experiment, all questionnaires were analysed by a team of social psychologists. They extracted, for each participant, a set of 17 numeric indexes summarizing participants' prejudice. More specifically, these data represent the psychological ground truth showing the participants' level of prejudice towards the black collaborator compared to the white one (at least, they provide insights into the levels of prejudice from a psychological

Table 1: Set of numeric indexes summarising the bias of the participant towards prejudice. These data represent the psychological ground truth showing the participant's level of prejudice towards the black collaborator. More specifically, indexes 6-9 and 14-17 represent the differential (between black and white people) version of indexes 2-5 and 10-13 respectively.

ID	Q. ID	Source	Description
1	IAT	Participant	Implicit prejudice. High score means a negative attitude towards black people.
2 - 5	Q1	Participant	Attitude towards black people. High score means a positive attitude towards black people.
6 - 9	Q1(D)	Participant	Differential attitude towards white and black people. High score means a better attitude towards white people compared to black people.
10	Q2	Participant	Score of the interaction with the black collaborator. High score means positive interaction.
11	Q2	Participant	Stress felt during the interaction with the black collaborator. High score means high stress.
12	Q2	Black coll.	Score of the interaction with the participant. High score means positive interaction.
13	Q2	Black coll.	Stress felt by the participant according to the black collaborator. High score means the black collaborator thought the participant was stressed.
14	Q2(D)	Participant	Differential score of the interactions with the white and black collaborators. High score means a better interaction with the white collaborator.
15	Q2(D)	Participant	Differential stress felt during the interactions with the white and black collaborators. High score means more stress felt during the interaction with the white collaborator.
16	Q2(D)	Both coll.	Differential score of the interaction with the participant. High score means the white collaborator judged the interaction better than the black one.
17	Q2(D)	Both coll.	Differential stress felt by the participant according to both collaborators. High score means more stress of the participant perceived by the white collaborator than the black one.

point of view). The 17 indexes are summarized in Table 1. Furthermore, the same psychologists manually inspected all the recorded videos in order to set the starting and ending point of each interaction: in this way the starting and ending parts of the video, in which the researchers are still present in the recording room, have been filtered to avoid eventual biases.

4. NONVERBAL BEHAVIORAL FEATURES

Each participant interacted both with the white and the black collaborators totaling 12 minutes of data (3 minutes on 2 topics with both the black and the white collaborators). Each recorded interaction produced a fairly significant amount of data recorded using different sensors operating at different sampling rates. More specifically: *(i)* RGB frames and the spatial coordinates each joint of the detected skeletons sampled at 10 Hz (Microsoft Kinect V2); *(ii)* Video and audio of the whole interaction, sampled at 30 FPS (GoPro camera); *(iii)* Estimated heart rate and galvanic skin response of the participant collected at 50 Hz (Shimmer GSR).

After having normalized the collected data by temporally aligning the three signals, we extracted spatial, audio, and biometric features from the Kinect, the GoPro and the Shimmer sensors respectively. The extracted features are listed below:

ID	Source	Description
1	Kinect	Mutual distance between interlocutors.
2	Kinect	Space (volume) between interlocutors.
3	Kinect	Movements of the upper body (participant).
4	Kinect	Movements of the center body (participant).
5	Kinect	Movements of the lower body (participant).
6	GoPro	Percentage of silence during dialogues.
7	Shim.	PPG biometric feature (participant).
8	Shim.	GSR biometric feature (participant).

It's worth emphasising that, for each participant, each feature has been computed in four cases: salient and non-salient interactions with both the white and black collaborators. In the rest of the paper, given a generic feature F, we will refer as F^{ws}, F^{wn}, F^{bs}, and F^{bn} to the features extracted in the four cases respectively. In rest of this section, the extracted features are better detailed.

4.0.1 Mutual distance

As mentioned above, for each tracked person, we used a cloud of 25 (x, y, z) points identifying the position of her joints in the space. In Figure 2 (left) are shown the actual joints tracked by the Kinect we used.

In each frame f in which both people are tracked, we can calculate the mutual distance of the two interlocutors as:

$$D_m(f) = dist(C^p, C^c) \quad (1)$$

where C^p and C^c are the centroids of participant and collaborator respectively. The two centroids are defined as the

center of mass of the joints of two interlocutors as follows:

$$C^p(f) = \frac{1}{m}\sum_{i=1}^{m} joint_i^p(f) \quad (2)$$

$$C^c(f) = \frac{1}{m}\sum_{i=1}^{m} joint_i^c(f) \quad (3)$$

where $m = 25$ is the number of joints tracked by the Kinect, $j_i^p(f)$ is the (x, y, z) triplet of coordinates of participant's joint i in frame f, $j_i^c(f)$ is the (x, y, z) triplet of coordinates of collaborator's joint i in frame f.

Once calculated, the mutual distance $D_m(f)$ can be either averaged over time windows of custom length, or over the whole interaction to get a coarse measure of how much the two interlocutors were close while talking. In this study we follow the second approach, so we take as feature the mean value over the entire interaction.

4.0.2 Space (Volume) between interlocutors

In order to capture with a single feature both the mutual distance between interlocutors and participant's gestures, we defined a novel feature that proved to be highly correlated with several psychological indexes and thus significant in detecting nonverbal behaviours.

For each frame of each interaction, we consider the cloud $P(f)$ of 26 (x, y, z) spatial coordinates representing the participant's joints and the collaborator's centroid (calculated as in equation 3). Then, for each frame, we use the Delaunay triangulation $S(f) = DT(P(f))$ to find a surface that passes through all the points in $P(f)$. Even though many different triangulations exist, the Delaunay's one is the most widely used. The volume contained within this surface is used as a feature. More formally:

$$F_{vol}(f) = Vol(DT(P(f))) \quad (4)$$

As shown in Figure 2, the base and the height of the cone are influenced by both the participant's movements and her distance from the collaborator respectively. Thus, this innovative feature is able to capture two important aspects of nonverbal behaviours at the same time. It's also worth mentioning that, despite its relevance, this feature cannot be accurately measured by a human coder but only using automated approaches.

4.0.3 Participant's movements

In order to capture the way participant's movements during the interaction we calculate three different measures, namely the quantity, velocity and acceleration. For this purpose we consider the upper, central, lower body joints as three separate sets. Once called J_{sel} the set of joints we want to consider, for each frame f in which participant's skeleton is tracked on the scene, we can compute:

$$m_i(f) = dist(joint_i(f-1), joint_i(f)) \quad (5)$$

$$\forall joint_i \ in \ J_{sel}$$

In this way we can calculate the total amount of movement that took place in frame f as follows:

$$M_p(f) = \sum m_i(f) \quad (6)$$

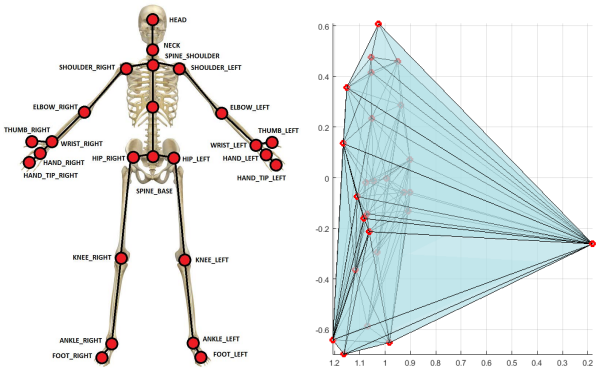

Figure 2: Representation of the 25 joints tracked by Microsoft Kinect V2 (left). Representation of the volume feature computed in one instant of the interaction. It comprises the volume between the participant's joints and the collaborator's centroid (right).

From this measure we then can recursively compute the velocity and acceleration of participant gestures in frame f as:

$$V_p(f) = \Delta(M_p(f), M_p(f-1)) \quad (7)$$

and

$$A_p(f) = \Delta(V_p(f), V_p(f-1)) \quad (8)$$

4.0.4 Pauses in dialogue

Regarding the pauses in dialogue, we compute a coarse-grained measure summarising the whole interaction. We start with a noise removal to obtain a cleaner signal. Considering the wide diversity of noise profiles, to achieve best results, this operation has been manually performed with the open-source software Audacity [1]. Then, given a cleaner signal y, we count the number of samples in which $|y| < thresh$. The threshold $thresh$ is set empirically through trial-and-error. In this way we have a measure of the percentage of silence taking place during the interaction.

4.0.5 Biometric features

As introduced above, we have two distinct biometric signals, namely PPG and GSR, related to heart rate and emotional arousal respectively. Both signals need a pre-processing phase to reduce noise. For this purpose, we apply to both signals a average filter of the form:

$$s_{filtered}(t) = \frac{1}{w}\sum_{t-w/2}^{t+w/2} s(t) \quad (9)$$

with $w = 10$.

Once obtained a clean signal we extract as features the signal's peaks to spot abrupt changes. At the time of writing, we noticed that compared with the duration of the whole interaction, PPG varies slowly while GSR extremely quickly. Due to this, we have found only subtle traces of correlation with psychological scores.

5. EXPERIMENTAL RESULTS

In this section we outline the most salient results achieved in this study. The first part discusses the degree of correlation found between extracted features and psychological scores. The second part, instead, shows how it is possible to infer psychological scores (i.e., IAT) using only automatically collected data (using a relatively simple classifier). These results are in line with research on the determinants of prejudice and, more specifically, with the idea that implicit rather than explicit prejudice is an especially relevant predictor of nonverbal behavior [10]. The fact that our results replicate the effects found in similar studies where nonverbal behavior was rated by external coders add to the external validity of findings. Moreover, identifying interpersonal distance and space volume as the two indices more related to implicit prejudice considerably extends previous research, by starting to clarify which are the main aspect that may be affected by prejudice.

5.1 Prejudice features assessment

We started by correlating the indexes of both implicit and explicit prejudice (provided by the team of psychologists) to the features extracted automatically from the sensors. In this phase, the features extracted from the salient and non-salient interaction have been treated separately. We made use of Pearson correlation coefficient, which definition for two random variables X and Y is given in equation (10). Pearson coefficient outputs a measure of the linear correlation of the input variables, where 0 means no correlation and 1 and -1 mean total positive (direct) correlation and total negative (inverse) correlation respectively.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{10}$$

For the sake of filtering the most significant correlations we used the p-value. P-value represents the probability of obtaining the same observed result following the null hypothesis. In this case, it is the hypothesis of no correlation. A p-value smaller than a specified threshold (called significance level α) suggests that the correlation might be significant. In this study, we set the significance level to the generally accepted value of $\alpha = 0.05$ [10].

Despite the noise levels within our data, several significant correlations (in both salient and non-salient interactions) emerged. In the following, we present the main trends we spotted within the obtained data and a possible explanation for each of them. All the significant correlations are summarized in Table 2 and divided into four main categories detailed below.

5.1.1 Influence of mutual distance

A positive correlation between the distance the participant keeps from the black collaborator and her implicit prejudice of participant (IAT score) actually exists. In particular, the distance the participant keeps from the collaborator is correlated to the IAT score through ($\rho = 0.43, pval = 0.02$) and ($\rho = 0.38, pval = 0.03$) in the salient and non-salient conversations respectively as shown in Figure 3 (top). Instead, there is no evident correlation between the mutual distance kept from the white collaborator and IAT score. In other words, participants with a higher implicit prejudice are likely to keep a smaller distance from white collaborators than from black ones.

Consistently, this correlation is even stronger if we consider the difference between the distance the same participant kept from the black and the white collaborator. Indeed, the features ($D^{bs} - D^{ws}$) and ($D^{bn} - D^{wn}$) correlate with the IAT score through ($\rho = 0.45, pval = 0.01$) and ($\rho = 0.43, pval = 0.01$) respectively. In other words, participants with higher implicit prejudice towards black people behave differently with the black and the white collaborator, keeping a larger distance with the latter.

5.1.2 Influence of space (volume) between interlocutors

Considering that the setting of the experiment was almost unconstrained (both interlocutors were free to move in the scene) the correlation obtained between IAT score and distance kept from black people is significant.

Nevertheless, by observing the space (volume) between interlocutors, we discovered another remarkable correlation. In fact, the correlation between the volume feature and the IAT score considering the interactions with black collaborators is characterized by ($\rho = 0.39, pval = 0.02$) in both salient and non-salient interactions as shown in Figure 3 (bottom). Also in this case, considering the difference between black and white collaborators leads to bigger relation with the participant's implicit prejudice. Indeed, ($F^{bs}_{vol} - F^{ws}_{vol}$) and ($F^{bn}_{vol} - F^{wn}_{vol}$) are related through IAT score through ($\rho = 0.47, pval = 0.006$) and ($\rho = 0.40, pval = 0.02$) respectively. Again, there's no correlation neither between the volume feature and the IAT score during interactions with whites. This seems to suggest that many participants tried to hide their actual level of prejudice when filling the first questionnaire, or that only their implicit prejudice leaked out in actual (nonverbal) behaviour.

Furthermore, we found another relevant correlation between F^{bn}_{vol} and Index 11 capturing the stress felt by the black collaborator ($\rho = 0.48, pval = 0.005$). Consistently, we find also a correlation between F^{bn}_{vol} and Index 17 which captures the difference between stress perceived by white collaborator and stress perceived by black collaborator ($\rho = -0.39, pval = 0.03$). These latter findings suggest two main hypothesis: firstly, that a participant with an high prejudice can disguise his real thoughts while filling the initial questionnaire, but his bias is usually not perceived by the black collaborator who is more likely to assign a lower appreciation to the interaction. Secondly, also these connections can be automatically captured by analyzing the mutual distance and space between the interlocutors.

5.1.3 Influence of motion during interaction

Various correlations suggest that the amount of motion taking place during the dialogue is related to the comfort of interlocutors and to the appreciation of the interaction. The results we obtained showed that hands joints movements contain the most of this information (see Table 2).

These results might have a twofold explanation: the former, related to cultural factors, might be related to the local heritage of moving hands for the sake of communication. The higher the level of comfort, the more various gestures are used for interacting. The latter, more psychological, might suggest that participants with higher prejudice levels towards black people are likely to freeze with the black collaborators.

5.1.4 Influence of pauses in dialogue

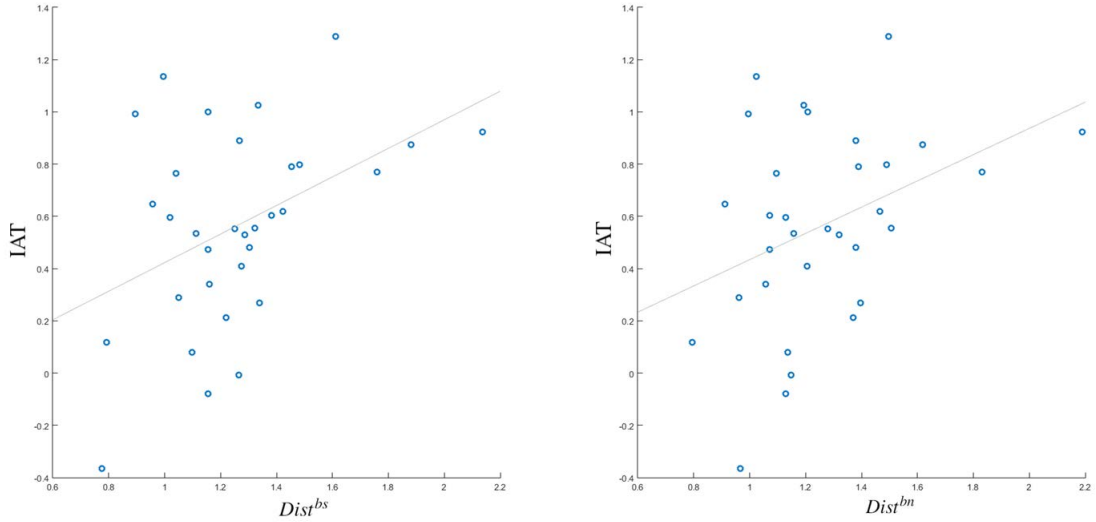

Figure 3: Scatter plots showing the linear correlation between the mutual distance and IAT score during both salient and nonsalient dialogues.

The presence and length of pauses during dialogues also seems to correlate with the appreciation of the interaction. In particular, we observe that the IAT score of the participant presents an inverse correlation with the percentage of silence during the non-salient conversation with the black collaborator ($\rho = -0.34, pval = 0.04$). We also found that the percentage of silence during the salient interaction with the black collaborator S^{bs} is directly correlated to the black collaborator's appreciation of the interaction (Index 10) through ($\rho = 0.41, pval = 0.01$). Consistently, we found also a direct correlation ($\rho = 0.37, pval = 0.03$) between silence percentage during the non-salient interaction with the black collaborator and Index 17 expressing the difference of stress felt by the participant according to the white and black collaborators.

At first, the fact that relaxed participants are likely to speak less during interactions may sound bizarre. Nevertheless, we can assume that a participant feeling discomfort might speak more to fill lapses in the dialogue. Indeed, speaking might be considered as a strategy to reduce stress levels. These results are coherent with the proved relation between the use of filler sounds (such as "ah" or uh") to fill lapses in speech and the shame or deceptiveness of the speaker, that has been extensively studied in psychology [23] [3].

5.1.5 *Influence of biometric features*

As far as biometric features are concerned, no significant correlation emerged neither with the IAT score nor with other questionnaire indexes. It is worth mentioning that, compared with the duration of the whole interaction, PPG varies slowly while GSR extremely quickly. Furthermore, it is worth mentioning that, during the study, the Shimmer device was paired with a workstation via bluetooth. Unfortunately, we often experienced failure of the bluetooth data streaming, which more than once led to incomplete data. Arguably, the missing data may have reduced several correlations' significance.

5.2 Prejudice Classification

Once determined the most significant features and their correlations with psychological indexes, we wanted to understand if it was possible to infer IAT scores using only automatically collected data. The pool of participants has been split in two clusters according to their IAT score. To do so, we labeled as "positive examples" and "negative examples' participants above and below the median of the IAT range. In this way, we reduced the problem to a 2 classes classification problem. This coarse simplification is justified by the fact that the IAT outputs pure numbers, significant only in the sample in which they were measured. From this standpoint, we can assume that the upper median is populated by participants with the highest prejudice levels in our sample. Our goal was to automatically separate participants that showed high prejudice levels from participant that showed low prejudice levels.

Because of the relatively small size of the dataset, we extended it for finding relations among the data. For each participant, we selected the features that exhibited a significant correlation (positive/negative) with the IAT score (visible in Table 2). Moreover, we formed couples of items by concatenating each feature vector with the feature vector of every other participant. The resulting vector was labeled with 1 if at least one of the two original vectors was labeled as "positive example", 0 otherwise. Formally, considering the original dataset composed of m examples, the size of the extended dataset (couples without repetitions) is given by:

$$\binom{m}{k} = \frac{m!}{k!(m-k)!} \tag{11}$$

With $m = 32$, $k = 2$, we obtain 496 examples.

We performed two different evaluations in order to asses the capability of our solution to deal with a complete dataset and eventually generalise to unseen element.

Firstly, (*Shuffle+Split* see Table 3) we trained the classifier

Table 2: Summary table of the most significant correlations we have found, showing p-values above $\alpha = 0.05$

Feature	Index	Rho	P-Value
$Dist^{bs}$	1 (IAT)	0.43	0.02
$Dist^{bn}$	1 (IAT)	0.38	0.03
$Dist^{bs}$-$Dist^{ws}$	1 (IAT)	0.45	0.01
$Dist^{bn}$-$Dist^{wn}$	1 (IAT)	0.43	0.01
F_{vol}^{bs}	1 (IAT)	0.39	0.02
F_{vol}^{bn}	1 (IAT)	0.39	0.02
F_{vol}^{bs}-F_{vol}^{ws}	1 (IAT)	0.47	0.006
F_{vol}^{bn}-F_{vol}^{wn}	1 (IAT)	0.40	0.02
F_{vol}^{bn}	11	0.48	0.005
F_{vol}^{bn}	17	-0.39	0.03
Mot_{hand}^{bs}	5	0.36	0.03
Vel_{hand}^{wn}	6	0.37	0.03
Vel_{hand}^{bs}	2	0.37	0.03
Vel_{hand}^{bs}	4	0.36	0.03
Vel_{hand}^{bn}	5	0.33	0.05
Vel_{hand}^{bn}	9	-0.38	0.02
Acc_{hand}^{wn}	6	0.40	0.02
Acc_{hand}^{bs}	2	0.35	0.04
Acc_{hand}^{bs}	4	0.35	0.04
Acc_{hand}^{bn}	5	0.34	0.04
Acc_{hand}^{bs}	10	0.35	0.04
Acc_{hand}^{bn}	10	-0.39	0.02
Vel_{ankle}^{bs}	12	0.34	0.04
Acc_{ankle}^{bs}	12	0.34	0.04
$Silence^{bn}$	1 (IAT)	-0.34	0.04
$Silence^{bs}$	10	0.41	0.01
$Silence^{bn}$	17	0.37	0.03

by shuffling and splitting the extended dataset into training and test sets respectively. We retained the 70% of elements as training set and used the remaining ones as test set. For classification, we made use of Adaboost with a number of trees fixed to 1000.

Secondly, (*Leave-One-Out* see Table 3) we performed a leave-one-out (LOO) validation as follows: we left one participant out, then we augmented the remaining dataset as described above and finally we concatenated the left-out participant's feature vector with itself and predicted his/her IAT score (*Leave-One-Out IAT* in Table 3). This experiment has been conducted using only the features showing a significant correlation with the IAT. Nevertheless, following the LOO validation scheme, we tested it also using the whole feature set (*Leave-One-Out ALL* in Table 3). For each one of the three experiments, we computed the $F_1 score$ as a measure of the classification's goodness and additionally reported precision and recall.

Table 3 shows, as expected, that using the whole dataset

Table 3: Table summarizing classification results in different settings.

Procedure	Features	Precision	Recall	F1
Shuffle+Split	IAT	0.93	1	0.96
Leave-One-Out	IAT	0.73	0.94	0.82
Leave-One-Out	all	0.54	0.76	0.63

(*Shuffle+Split*) leads to the highest classification accuracy. However, test subjects have been used during the training stage eventually introducing classification artefacts. Conversely, the LOO test measures the ability of classifying unseen subjects. Even in this setting (*Leave-One-Out IAT*), we still achieved an $F1$ score around 82%. Furthermore, Figure 4 showing a detailed view of our classification results, leads to interesting observations. In particular, all classification errors are concentrated in the lower section of the IAT spectrum. We think that it is possible that higher IAT values might lead to more evident (and maybe more structured) body movements while lower IAT values can be expressed in more subtle and subjective ways. It is also worth mentioning that the reported results have been obtained without features normalisation. Indeed, we empirically found that normalisation led to slightly worse accuracy levels. Finally, the lowest score (*Leave-One-Out all*) has been achieved by making use of all the available features. This result empirically corroborates our psychological conclusions showing that features poorly correlated with the IAT affect negatively the prediction. This effect should be further investigated to verify if the cause resides in the subjectivity of those features.

Although being aware that these are pioneering studies and that more sophisticated machine learning's algorithms could be applied, we find these results - introducing how to infer people prejudice in a completely automatic fashion - relevant for the field.

6. CONCLUSION

In this paper we presented a study on identifying prejudice with nonverbal behaviours. More specifically, an experiment about prejudice on 32 individuals has been performed and the collected data have been used to automatically infer the score of IAT tests. To the best of our knowledge, this is the first and most detailed study on the automatic detection of INVB so far. We demonstrated, for the intergroup case, the relevance of shifting the analysis of psychological experiments from subjective evaluations of human coders to objective measures. This paradigmatic change, in fact, could lead to extremely cheaper and eventually more accurate analysis. The analysed features were partly derived from previous psychological research, that identified indices such as interpersonal distance, body posture and movements as especially relevant indicators of prejudice [17]. Other relevant aspects, such as eye gaze, could not be assessed, due to the procedure used (we would have needed a camera placed just in front of the participant, undermining his/her possibility to move around during the interaction and, thus, of assessing the other indices). Furthermore, the space volume index is really kind of a combination of previous indices, allowed by the new technology and, to our knowledge, never explicitly hypothesised in previous psychological research on intergroup relations and prejudice.

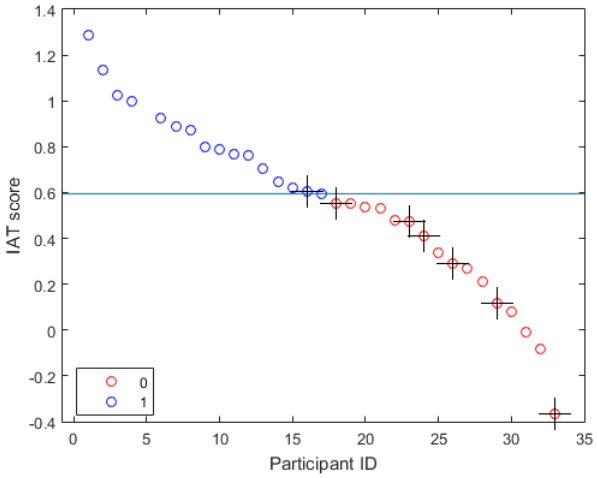

Figure 4: Detailed representation of leave-one-out classification result showing participants sorted according by their IAT score. The reference line is the median IAT score eventually separating subjects with high and low prejudice levels. Wrongly classified participants are highlighted with black crosses. The figure shows how classification errors are actually concentrated within the lower end of the IAT spectrum. This might suggest that higher IAT levels might lead to more explicit effects on body movements.

Although the promising results we achieved, several limitation still separate this work from practical applications. First of all, current instrumental accuracy still require controlled conditions to perform experiments. Secondly we still have to inquire whether computational models derived from a group of subjects could be actually used to classify a different group of subjects (e.g., detect prejudice directed to different communities). Nevertheless, despite these initial limitations, it is also worth noticing that despite this study has been tailored on prejudice, the same approach and set of technologies could be used in a large variety of application domains. For example, it could be used in schools to identify children with anxiety issues (i.e., ADHD), in self-driving cars to assess driver attention level, in research to automate the annotation of psychological experiments, until border control to spot possibly dangerous behavioural outliers.

7. ACKNOWLEDGMENTS

This work was carried out within the project *La Città educante* (ctn01_00034_393801) of the National Technological Cluster on Smart Communities cofunded by the Italian Ministry of Education, University and Research (MIUR).

8. REFERENCES

[1] Audacity Team. 1999-2016. Audacity. (1999-2016). http://www.audacityteam.org/

[2] Tobias Baur, Ionut Damian, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2013. *Human Behavior Understanding: 4th International Workshop, HBU 2013, Barcelona, Spain, October 22, 2013. Proceedings*. Springer International Publishing, Cham, Chapter NovA: Automated Analysis of Nonverbal Signals in Social Interactions, 160–171. DOI: http://dx.doi.org/10.1007/978-3-319-02714-2_14

[3] Stefan Benus, Frank Enos, Julia Bell Hirschberg, and Elizabeth Shriberg. 2006. Pauses in deceptive speech. Proc. ISCA 3rd International Conference on Speech Prosody.

[4] A. Bharathi, A. Singh, C. Tucker, and H. Nembhard. 2016. Knowledge Discovery of Game Design Features By Mining User-Generated Feedback. *Computers in Human Behavior* 60 (2016), 361–371.

[5] Nicola Bicocchi, Damiano Fontana, and Franco Zambonelli. 2014. A self-aware, reconfigurable architecture for context awareness. In *IEEE Symposium on Computers and Communications, ISCC 2014, Funchal, Madeira, Portugal, June 23-26, 2014*. 1–7.

[6] Nicola Bicocchi, Matteo Lasagni, and Franco Zambonelli. 2012. Bridging vision and commonsense for multimodal situation recognition in pervasive systems. In *2012 IEEE International Conference on Pervasive Computing and Communications, Lugano, Switzerland, March 19-23, 2012*. 48–56.

[7] Keith Brawner and Benjamin Goldberg. 2012. Real-time monitoring of ecg and gsr signals during computer-based training. In *Intelligent Tutoring Systems*. Springer, 72–77.

[8] Nathan Burba, Mark Bolas, David M Krum, Evan Suma, and others. 2012. Unobtrusive measurement of subtle nonverbal behaviors with the Microsoft Kinect. In *Virtual Reality Short Papers and Posters (VRW), 2012 IEEE*. IEEE, 1–4.

[9] A. Burns, B.R. Greene, M.J. McGrath, T.J. O'Shea, B. Kuris, S.M. Ayer, F. Stroiescu, and V. Cionca. 2010. SHIMMER x2122; x2013; A Wireless Sensor Platform for Noninvasive Biomedical Research. *Sensors Journal, IEEE* 10, 9 (Sept 2010), 1527–1534. DOI:http://dx.doi.org/10.1109/JSEN.2010.2045498

[10] John F Dovidio, Kerry Kawakami, and Samuel L Gaertner. 2002. Implicit and explicit prejudice and interracial interaction. *Journal of personality and social psychology* 82, 1 (2002), 62.

[11] Denise Frauendorfer, Marianne Schmid Mast, Laurent Nguyen, and Daniel Gatica-Perez. 2014. Nonverbal Social Sensing in Action: Unobtrusive Recording and Extracting of Nonverbal Behavior in Social Interactions Illustrated with a Research Example. *Journal of Nonverbal Behavior* 38, 2 (2014), 231–245. DOI:http://dx.doi.org/10.1007/s10919-014-0173-5

[12] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, and others. 2014. The Distress Analysis Interview Corpus of human and computer interviews.. In *LREC*. 3123–3128.

[13] A. G. Greenwald, D. E. McGhee, and J. L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology* 74 (1998), 1464–1480.

[14] A. G. Greenwald, B. A. Nosek, and M. R. Banaji. 2003. Understanding and using the implicit association

test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology* 85 (2003), 197–206.

[15] S. Handri, K. Yajima, S. Nomura, N. Ogawa, Y. Kurosawa, and Y. Fukumura. 2010. Evaluation of Student's Physiological Response Towards E-Learning Courses Material by Using GSR Sensor. In *Computer and Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on*. 805–810. DOI: http://dx.doi.org/10.1109/ICIS.2010.92

[16] Béatrice S Hasler, Oren Salomon, Peleg Tuchman, Ady Nae O'Malley, and Doron A Friedman. 2011. Real-time Translation of Nonverbal Communication in Cross-Cultural Online Encounters. In *Proc. of CMVC*.

[17] M.R. Hebl and Dovidio J.F. 2005. Promoting the social in the examination of social stigmas. *Journal of personality and social psychology* 9, 9 (2005), 156–182.

[18] Jin Joo Lee, Brad Knox, and Cynthia Breazeal. 2013. Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions. (2013).

[19] D. Matsumoto and H. C. Hwang. 2016. *The cultural bases of nonverbal communication*. 77–101.

[20] Dimitris Metaxas and Shaoting Zhang. 2013. A review of motion analysis methods for human nonverbal communication computing. *Image and Vision Computing* 31, 6 (2013), 421–433.

[21] Joann M. Montepare. 2014. Nonverbal Behavior in the Digital Age: Meanings, Models, and Methods. *Journal of Nonverbal Behavior* 38, 3 (2014), 279–281. DOI: http://dx.doi.org/10.1007/s10919-014-0187-z

[22] M. L. Patterson. 1982. A sequential functional model of nonverbal exchange. *Psychological Review* 89 (1982), 231–249.

[23] Suzanne M Retzinger. 1995. Identifying shame and anger in discourse. *The American Behavioral Scientist* 38, 8 (1995), 1104.

[24] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency. 2013. Automatic Nonverbal Behavior Indicators of Depression and PTSD: Exploring Gender Differences. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. 147–152. DOI:http://dx.doi.org/10.1109/ACII.2013.31

[25] Grace A Sullivan and Wind Goodfriend. 2013. The Effects of Controlling Nonverbal Intimacy. *Journal Contents* 18, 2 (2013), 32–41.

[26] Andrea Stevenson Won, Jeremy N Bailenson, and Joris H Janssen. 2014a. Automatic detection of nonverbal behavior predicts learning in dyadic interactions. *Affective Computing, IEEE Transactions on* 5, 2 (2014), 112–125.

[27] Andrea Stevenson Won, Jeremy N Bailenson, Suzanne C Stathatos, and Wenqing Dai. 2014b. Automatically detected nonverbal behavior predicts creativity in collaborating dyads. *Journal of Nonverbal Behavior* 38, 3 (2014), 389–408.

[28] Yang Rui Zhang, Guang Huo, Jian Feng Wu, Jun Bo Yang, and Chen Pang. 2015. An Interactive Oral Training Platform Based on Kinect for EFL Learning. In *Applied Mechanics and Materials*, Vol. 704. Trans Tech Publ, 419–423.

[29] Zhengyou Zhang. 2012. Microsoft kinect sensor and its effect. *MultiMedia, IEEE* 19, 2 (2012), 4–10.