Understanding Social Signals from Nonverbal Behaviors in a Mobile Setting

Xin Jia

MSC. THESIS

Human Media Interaction, University of Twente
2017, August

GRADUATION COMMITTEE

Dr. ing. G. Englebienne
Dr. A. Bulling (Max Planck Institute for Informatics)
Dr. M. Poel

Abstract

Nonverbal behaviors are natural yet critical channels in understanding social signals. The automation of the apprehension of such signals has been an increasingly popular topic in recent decades, due to the development of the recording hardware as well as the machine analysis capabilities.

In this study, predictive models for measuring the emotions, attitudes and personalities of individuals from nonverbal behaviors were established. The realization of the model involved the construction of a multimodal and mobile recording framework of behaviors, the collection of individuals' emotions, attitudes and personalities as ground truths, and the application of various machine learning algorithms which find and interpret the patterns in the data.

A user study was designed in order to obtain the necessary visual, audio and spatial data. 20 participants were recruited and requested to have dyadic conversations with pedestrians on the street. The conversations were recorded and then processed in order to extract the following features: facial expression, gaze location, interpersonal distance, speech data and so forth. Furthermore, the participants reported their experiences after each conversation, including the perceived friendliness of the pedestrian, the levels of frustration after the conversation, as well as their emotion states in the arousal-valence mode. Finally, the participants completed a set of psychological questionnaires regarding their personality and racial prejudice level at the end of the whole experiment.

With the extracted nonverbal behaviors as features and the results from questionnaires as the ground truth, models have been trained to predict the above measurements.

Acknowledgements

I would like to express my great appreciation to my external supervisor Andreas Bulling, who has been much more than a supervisor during my stay in Max Planck Institute for Informatics in Germany. His support, active involvement and strict criteria for science research have motivated me throughout this project. I would also like to thank my daily supervisor Philipp Müller, who has always been so generous in giving his guidance and help in every details of the project. Without his knowledge in study design, psychology and data analysis, it would not have been possible.

I would also like to thank my internal supervisor Gwenn Englebienne. During the remote meetings every week or every other week, he has provided so much insights and suggestions regarding the experiment design, user study ethics and data analysis. Thanks to his active involvement in the project, I was able to discuss with him and think over the every details of my project.

I am also grateful to Michael Xuelin Huang, who has been so kind in giving suggestions, referring to useful previous work and literature, and summarising the improvements I could work on. I also appreciate Julian Steil for his help in gathering and installing the devices and introducing eye tracking technologies and theories to me. Likewise, I would like to give thanks to Xucong Zhang for his selfless instruction during the hardware framework construction stage and his sharing of gaze estimation models.

Special thanks to all my participants in my user studies, who have been patiently wearing the recording devices, trying to follow the requirements of each step of the study, and bravely having conversations with strangers. Without their cooperation, it will be even impossible for me to collect the data I need to train on.

I would also like to thank Max Planck Institute for Informatics, as an institution that supports and funds my project work, and providing the access to experiment devices, academic resources and technical assistance.

I would like to thank all the people I have met in Max Planck Institute for Informatics, for setting a role as an independent, energetic, knowledgeable and motivated researcher.

Finally, I want to thank my family and my friends, for listening to me, comforting me, and encouraging me throughout the project.

Contents

Chapter 1 Introduction

Extracting social signals from nonverbal behaviors has been attracting more and more attention from the academia in recent decades. Unlike verbal signals conveyed by the content of speech, nonverbal behaviors have been proved to be require high cognitive capability, therefore are not easily observed and manipulated by the people themselves.

The social signals can be the social actions, social emotions, social interaction, social attitudes, social traits and social relationships [1], depending on the time scale. In this study, we have chosen the several topics in the field of social emotions, social interactions and social attitudes, including the emotional states of individuals, the level of friendliness an individual gives during conversation, the personality and the level of racial prejudice of individuals. Standard psychological methods and supplementary self-report questionnaires were adopted in the study to evaluate the above mentioned social signals.

The goal of the study is to spot and extract the important nonverbal behaviors automatically, and relate such features with the social signals as evaluated by the standard psychological measurements. In order to build the model that could predict those measures from the nonverbal behaviors, a study was designed to connect the nonverbal behavior and the psychological measures. There are two phases of the study: collecting data set of nonverbal behaviors and the corresponding measures; and implement machine learning algorithms to represent and estimate their relationships.

There are challenges in developing multimodal framework to capture and record nonverbal behaviors. First of all, it has to make sure all the channels from different devices were correctly gathering data, and such data should be synchronized for future processing. Second, the processing of the multimodal data involves knowledge in various fields, including speech analysis, computer vision and so forth, therefore requires work in diverse directions. This also means multiple device and recording software for data collection, multiple software or program to preprocess raw data, and multiple programs to extract features. Finally, it is appreciated to not only represent the dynamics of data from each channel, but also the interactive effect between channels.

Furthermore, the study is carried out in a mobile approach, which introduces two major difficulties: much more noise in the raw data collected; increase the randomness of the experiments. Noise refers to the unwanted dynamics in the raw data, such as too strong sunlight, which could increase the difficulty for eye movement analysis, or too loud noise, which invites noise into the speech analysis. The randomness refers to the fact the when the data collection is in the field without the guidance from the experimenter, participants will to some extend deviate from the procedure or requirements as defined by the experimenter.

To avoid or diminish the possible source of errors, multiple iterations of designs were carried out.

In chapter 2, established theories and previous work in the field of social signal processing, emotional state, personality and racial prejudice will be reviewed and summarized.

Chapter 3 describes the methodologies that have been used in the study. The major changes in the iterations of design will be listed and explained, followed by a detailed description about the final version of study design and design of experiment procedures.

Chapter 4 explains about the set-up and procedures for data collection, including the explanation about apparatus, software for recording used in the study, and the actual procedure of data acquisition.

In chapter 5, the details of the data analysis will be given. This includes the depiction of preprocessing of raw recording data, processing of psychological measurements, actions in feature extraction, as well as the pipeline of model training. A mind map is introduced to simplify the pipeline.

Chapter 6 exhibit the training process and performance of various models following the mind map described in chapter 5.

Chapter 7 summarizes the obstacles met throughout the project, both in the user study stage and the data mining stage. Future work is also discussed about in this chapter.

Chapter 2 Related Work

2.1 Racial Prejudice

For almost 100 years, the study of racial and national prejudice has been gaining increasing attention from the researchers, the government and the society. Prejudice is defined to be a negative attitude toward a group or toward its members, while stereotypes are the mental concept of the groups in question [2]. The direct social and health impacts of social prejudice and discrimination of the ethnical minority members include the inferior economic situation, adverse effects on mental and physical conditions [3] and diminished access to opportunities [4].

Racial prejudice can be divided into two types: implicit and explicit prejudice [5]. Adopting the definition from [6], racial prejudice could further split into three: public, personal but conscious, and implicit. The public racial prejudice means attitude publicly shown by the user, which could be influenced by the social expectations and the will for impression management. The personal but conscious racial prejudice stands for the attitude an individual adopts that is not expressed publicly, but consciously aware of. The implicit racial prejudice represent the unconscious feelings and beliefs of the individual.

Owing to the historically rooted or contemporarily established social norms of egalitarianism which discourage the expression as well as personal acknowledgement of bias [7], the representation of racial prejudice has changed from explicit and blatant discrimination toward ethnic minorities, to implicit and subtle prejudice [8]. This has led to the phenomenon of modern racism, symbolic racism, ambivalent racism, aversive racism, laissez-faire racism and subtle racism [9].

The common methods to measure explicit racial prejudice include Bringham's Attitudes Toward Blacks Scale [10] and McConahay's Modern Racism scale [11], where both of the scales are specifically designed for black-white racial prejudice evaluation scenario. The methods for implicit racial prejudice are designed into more diverse forms. Among them, the implicit association test [12] is the most frequently adopted measure, which can be modified in order to suit various topics. Other measures are semantic priming technique [13], evaluative priming technique [14], word-completion task [15], and Go/No-Go Association Task [16], and so on.

Laboratory research [17] demonstrates that explicit prejudice measures are related to the ratings of individuals' verbal racial bias, while the nonverbal friendliness predicts the implicit racial prejudice better. Although researchers believe that both verbal and nonverbal behaviors partially reflect the true racial prejudice levels of individuals, it is also recognized that individuals might deliberately alter their utterances and behaviors due to self-presentational concerns [18]. The dual attitude model [19] states that when cognitive capacity and motivation are sufficient, people tend to regenerate their attitudes. Considering explicit attitudes in verbal representations, they are within full access to the person, therefore are easier to monitor and manipulate [20]. Nonverbal behaviors, however, as is reported in [21], lie outside of conscious awareness and control, and are prone to leak the individual's real attitude. Furthermore, according to the research from Dovidio et al., [22] rather than the intendedly altered verbal behaviors, the subtle nonverbal behaviors turn out to essentially determine the observers' perceived friendliness. Therefore, nonverbal behaviors can be and do serve as an effective means to measure an individual's racial prejudice.

Racial prejudice can be communicated in varied non-verbal forms. Utilizing the gaze data, researchers could predict an individual's racial prejudice by examining his/her behaviors in blinking rate [23], location of gaze fixation [24], and visual contact [25].The auditory nonverbal behavior such as the tone volume, pitch and intonation [26], speech latency [27], stuttering and laughing [28] also reflect the individual's attitude toward certain subject. Seating distance [29], orientation, posture, head nodding, facial expressions [30] are also considered to be correlated with an individual's level of racial prejudice. Other contributive measures include physiological and neural responses including cardiovascular responses [31], blood pressure measures and heart rate.

Palazzi et al. [32] have claimed to be the first group working on automatically measuring racial prejudice from nonverbal behaviors. In their research, a set of measures were taken in order to extract the systematic relations between nonverbal behaviors and racial prejudice. Nonverbal behaviors including mutual proximity, space between interlocutors, movement of different body parts, the percentage of silence during dialogues, and PPG and GSR biometric features were utilized in the model. The standardized methods including the questionnaires for explicit racial prejudice, as well as the implicit association test for implicit racial prejudice were also performed. Results turn out that the mutual distance, space volume between interlocutors and the motion during interaction are the most significant factors to indicate the individual's level of racial prejudice, especially the implicit measure. Furthermore, the trained model proved to have a precision score of 0.73 and F1 score of 0.82 when classifying whether an individual has high racial prejudice or low racial prejudice in a leave-one-out manner. However, besides the major drawback that the selection of features are based on the test data, the experiment design still suffers from several drawbacks. None of the auditory data, facial expressions and gaze data were implemented in the model, therefore failing to investigate about the possible contributions and strengths of each channel of input. Furthermore, the design of the process ignores the importance of controllable variable, such as the sequence of experiment and standard tests which might alter the behaviors of individuals when the individual realise the goal of the experiment before interacting with experimenters.

Although previous research directly studying the topic of racial prejudice measurement with automatic nonverbal behavior analysis remains very limited, the range and choice of data channels, the design of the experiments, and the means to process observational data from other works in emotion detection and attitude recognition bring inspirations to the multimodal recognition of racial prejudice.

2.2 Personality

Personality refers to an enduring personal characteristics that emerge consistently in behaviours in various situations [1]. It can be measured with diverse tests, including the Big Five Inventory [33], Minnesota Multiphasic Personality Inventory (MMPI-2) [34], Neurotic

[1] https://en.wikipedia.org/wiki/Personality

Personality Questionnaire KON-2006 [35] and so on. Besides the psychological measurements, nonverbal behaviours are also considered to emit signals which convey information about personalities of people [36].

Speech features, especially the prosodic features such as pitch, tempo and energy have been long deemed to be indicative of the personality of the speaker [37]. Early studies in speech analysis [38] has proved from experiments that the pitch and rate of speech considerably determines people’s impression of the speaker’s personality. Specifically, high-pitched voices were related with properties such as less truthful, less emphatic, and more nervous, while slow-talking speakers were regarded as less truthful, less persuasive and more passive. Psychologists have also shown that shorter silent and filled pauses, higher voice quality and intensity, higher pitch and higher variation of the fundamental frequency of the speech appear more in extravert individuals [39].

In the work of Polzehl et al., [40], a support vector machine model was established that generates the ratings using the NEO-FFI personality inventory from the recordings of one professional speaker, by using the prosodic and acoustic speech properties. Results turned out that the neuroticism and extraversion scores can be classified best, and high and low conscientiousness can be discriminated clearly, while openness can not be predicted from the speech features. Another work [41] has focused on the prediction of extraversion from the Big Five Inventory, and had adopted a multimodal framework for feature generation. The features include speech features such as formant frequency, energy in frame, length of voiced segments and unvoiced segments and so on, as well as visual features which indicate the intensity of motion of different body parts. The prediction performance (89.14%) was proved to be well above a baseline situation (66.7%) of always assigning the most frequent class to a new sample.

Gaze is also known as an important signal to reveal an individual’s personality. Gaze aversion is associated with passive traits such as shyness and emotional overcontrol [42]. During face-to-face conversations, a strong correlation was found between the activity of gazing at the interlocutor and the agreeableness score of the individual. Similarly, frequent mutual gaze is related to the sum of agreeableness of both of the speakers [43]. Furthermore, the curiosity level can also be predicted from the eye movements of individuals, according to the experiment results from Hoppe et al., [44].

Personal spatial zone is markedly influenced by the individual’s personality traits, as was concluded from an experiment in human-robot interaction [45]. In the experiment, those who maintain larger distance with a human-sized conversational robot tend to achieve higher proactiveness scores in personality measurements. Another study showed that interpersonal distance and orientation are determined by the personal characteristics, such as the warmth and dominance of the two speakers, together with other factors in social situation.

2.3 Emotion

Human express their emotions through verbal or nonverbal behaviours, either intuitively or unknowingly, and nonverbal behaviours are believed to be more closely related to the real emotion of the individual.

Generally speaking, emotions are usually defined in two kinds of systems: by basic discrete emotions, or by the two dimensional arousal-valence model [46]. The basic emotions consist of distinct physiological experiences: anger, disgust, fear, happiness, sadness and surprise, along with the subclasses under each category. The arousal-valence model, however, divide and assign all emotions into a two-dimensional space, where the valence represents how negative or positive the experience is, and arousal means energized or enervated the experience feels.

Facial expressions are widely validated as universal signals for emotion, regardless of ages, genders and cultures [47]. The facial expressions are generated by contractions of facial muscles, and there are two main methodological approaches to read facial expressions: the judgement-based approach, which focus on the emotional messages conveyed by facial expressions; and the sign-based approach, which can be regarded as the decomposition of basic facial expressions and does elementary coding of the facial motion and deformation into visual classes [48]. Compared to judgement-based facial expression approach, the sign-based approach is treasured due to its capability in detecting and representing slight differences, practicability in automation and universality. A standardized system was hence established: Facial Action Coding System, and each movement is categorized into specific Action Units (AUs), and labelled with presence and intensity.

Speech is viewed as a major channel for emotion expression. Besides the verbal expressions that are linguistically emotion-relevant, the nonverbal vocal expressions are also important clues of affection. Acoustic features such as pitch, energy, duration, rate, spectral energy and their functionals have been utilised in various studies [49, 50] to predict the emotional state of the person. In order to standardize and benefit the tedious work of feature extraction from speech, a minimalistic set of acoustic parameters [51] were recommended by a community of psychologists, linguists and computer scientists. This Geneva Minimalistic Acoustic Parameter Set (GeMAPS) was designed specifically for voice research and affective computing.

Emotions are conveyed also in other channels, for instance the eye movements or status of the person. In an experiment [52] where pupil diameter was monitored during picture viewing, the results have shown that pupillary changes were larger when viewing emotionally arousing pictures, and this led to the hypothesis that pupil’s responses reflect emotional arousals.

2.4 Tools and platforms

To capture and measure the nonverbal behaviors from different modalities, numerous frameworks were designed and built, attempting to simplify the procedure of data acquisition and make the processing more convenient and accessible. Such advancements in the toolbox makes the multimodal subtle signal acquisition and processing possible. Social Signal Interpretation (SSI) framework [53] brings forward a platform where the pipeline of recording, analysis and recognition can be realised in the same system. It supports streaming from multiple channels including audio, visual, motion and physiological signals. Another similar multimodal frameworks is EyesWeb [54]. On the other hand, Affectiva [55], Computer Expression Recognition Toolbox [56] and openFACE [57] are specific software focusing on the facial landmark detection, facial action classification and facial expression processing. The openFACE is capable of processing gaze data, whereas the PyGaze [58] is an open-source toolbox exclusively for eye tracking. Praat [59] is a tool specialized for the analysis of auditory data. Its functionalities are spectral, pitch, formant and intensity analysis, as well as annotation and manipulation of audio file. Furthermore, machine learning methods are readily embedded in the software and therefore allows for pattern mining from the speech data.

Researchers have also explored the possibility of extension into deep learning approaches in the data mining step of the experiment. Poria et al. [60] designed a framework which adopts deep convolutional neural networks for the feature extraction from visual and textual modalities, and multiple kernel learning classifier for emotion recognition. Kim et al. [61] instead proposed a model called Deep Belief Network model, where non-linear interactive audio-visual features can be extracted even in an unsupervised context. Results from the research indicates a potential of deep learning algorithms in the processing of multimodal data.

To enrich the toolbox for the study, Bousmalis and his colleagues [62] have listed the potentially useful cues and their automatic measurement tools in nonverbal behaviors of agreement and disagreement, which could also bring inspiration to the similar implicit attitude detection task of racial prejudice. In the work of Zeng et al. [63], a review of the automatic methods for facial and vocal affect recognition methods, particularly, the natural and spontaneous setting were summarised.

Chapter 3 Methodology

In this section, the design of the user study and the important iterations of changes in the design will be introduced.

3.1 Design Iterations

The study aims to investigate the relationship between people's nonverbal behaviours when talking with strangers and those people's emotions, attitudes or traits. Therefore, the key point for the study design is to capture and record as many channels of nonverbal behaviours as possible in an unobtrusive, accurate and natural way. Furthermore, an objective and sound measurement of the target group's emotions, attitudes and traits should be established. Hence, the experiment design should be tailored in order to meet these requirements.

In the early stage of the experiment planning, several attempts of designs were made but discarded after consideration, discussion or pilot study:

3.1.1 Physiological Measurements

Physiological measurements such as heart rate and galvanic skin response were at first considered as additional channels of social signal. However, according to a previous studies in the a similar setting about predicting racial prejudice [64], weak correlations were found between the biometric measures and the ground truth due to unavoidable noise. Furthermore, the request of wearing biometric devices makes it even harder to propose a feasible cover story to prevent the reveal of the real intention of the study. Lastly, on-body sensors might augment people's behaviours even more. Therefore, the above mentioned physiological measurements were discarded.

3.1.2 Multiracial Study

On the grounds that the majority of the previous studies about racial bias focus on the White-Black scenario, while the remaining group also fixate on two race scenarios, such as White-Arab or White-Asian, an attempt was made to include multiple races in the study. The preliminary plan was to request the participant to have conversations with people from different ethnic groups, including Caucasians, Arabs, Blacks and Asians. The reason for choosing Caucasians and Blacks is that this has been a standard setting in the racial prejudice related studies. The Arabs were chosen because in recent years, there are accumulating conflicts and gap between the Arabian/Muslim world and the western world, therefore Arabs/Muslims could also be a potential trigger of racial bias. Finally, Asians were chosen because previously, there were very few studies or little focus about the issue of racial prejudice toward Asians, and the western population are reported to have different types of racial prejudice toward Asians other than toward Blacks.

However, in the pilot study, several participants have shown different levels of learning effects in the racial implicit association test. The participants were required to take 3 implicit association tests in a row, so that we have an evaluation of their levels of racial bias toward Asians, Blacks and Arabs. The results turned out that the more implicit association tests the user has taken, the less bias was detected, and the responses seem to also speed up. Previous studies also back up this phenomenon [65]: participants who have previous experience of implicit association test on average achieve less significant results in terms of their racial bias, known as learning effect. Additionally, this effect is also influenced by the individual's motivation and ability to manipulate the score.

A first thought to solve this problem would be to adopt a balanced study, where people are divided into multiple groups, each group taking specific sequence of tests. However, there are two issues with such a solution: typically, a balanced study is for drawing a general conclusion about a population, not for evaluating the individuals; the learning effect, as explained before, heavily depends on the individual's ability and motivation to manipulate results, therefore can not be averaged across groups. As a result, we switched to a White-Arabs scenario, where the participant will only need to complete one implicit association test, in other words, the implicit racial bias test toward Arabs. The reason for choosing Arabs is that due to recent events, the racial difference or conflict with Arabs remains a prominent issue especially in Europe, making Arabs evident triggers for racial bias; furthermore, in the location of the study: Germany, there is higher presence of the Arab/Muslim group.

3.1.3 Controlled Setting

In the early stage of the project, the experiment was designed to replicate a similar situation as [66]'s work, where representatives of the two races will be recruited to act as participants, and the real participants will have conversations in separate rooms with their confederates. In such a setting, a set of devices suitable for fixed location were adopted. Kinect was designed to capture the color and depth information during the the conversations, therefore providing the distance between two people, and their body movements or gestures, as well as head pose. GoPros will be mounted on the chest to capture the facial expression and head movements of the two people. Additionally, headworn microphone records the utterance of two people. The advantages of this setting are that Kinect is able to do whole body tracking on both of the people, the GoPros are able to capture the facial expression and head movements of both speakers, therefore providing a more extensive feature set. All the participants will be conversing with the same representatives, therefore making sure the triggers for racial prejudice will be the same for every participant, making comparisons and general conclusions reasonable. Most importantly, since it is a controlled setting, all devices and the experiment processes will not be affected by the noise, weather, sunlight condition, too much dynamics or even other passengers on the street.

However, being able to draw general conclusions also requires that the triggers, or in other words, the representatives of the two groups should be a standard one. As there will be only one person for each ethnic group, selecting an impeccable sample becomes difficult. Besides appearance, height and clothing, other factors such as education background, income and personality can also be determining factors for the total impression the person gives. As a consequence, a random and large enough sample size for the representatives seem to be a better option. To facilitate this, the experiment has been changed to take place outdoors in public areas, and the participant freely chooses his or her confederate randomly. An additional advantage of the outdoor setting is that previously there has been little work done in nonverbal behaviour analysis in an egocentric view, making this a potentially beneficial topic to work on. Inevitably, the uncontrollable situation of the outdoor design also has much drawbacks, therefore precautions were taken as much as possible during the study design stage to remedy the disadvantages.

3.1.4 Task Content

In the early stage of the experiment, the task assigned for the participants was to discuss about a given topic with pedestrians. The reason for such a task was simply to justify the action to approach and converse with strangers. However, the strict requirements for participants and the long duration of the experiment turned out to be a difficult setting for recruitment. Attempts were made to recruit people by posters, flyers, group emails, face to face recruitment and so on, but the effects were limited. Therefore, we changed the task of the experiment to participant recruitment, in which the participant not only have conversations with different groups of people, but also assist to recruit new participants.

This decision is believed to be logical. First of all, the participants were given a task, therefore having a reason to approach strangers. Second, the requirement for recruiting people to some extent pushes the participants to interact with as many strangers as possible. Third, the success and failure during the recruitment process could influence the emotional states of the participants, therefore making the emotional states of the participants more widely distributed.

3.1.5 Ground Truth Expansion

During the early stage of the experiment design, the topic of the study was investigating the relationship between nonverbal behaviours in conversations with strangers from two ethnic groups and people's level of racial prejudice toward Arabs. However, as the experiment progressed, the participants admitted or the experimenter noticed from the recorded videos that the participants faced difficulties in finding or recognizing Arabs in the campus. Such a finding put the topic about racial prejudice into risks. Furthermore, obstacles in recruiting new participants made a large enough training set infeasible. As a result, other potentially useful topics were also introduced to the experiment, by requesting the participant to complete more related questionnaires and tests. The topics include: predicting personality from participants' nonverbal behaviours; predicting the emotional state, i.e. arousal and valence, or level frustration of participants after each conversation; predicting the perceived friendliness of pedestrians after each conversation.

3.2 Final Design

After attentive reasoning and discussion, the final plan went as follows: participants were equipped with a set of wearable devices such as color and depth cameras and a microphone and had conversations with random pedestrians on campus. After being told a cover story for the task, they were instructed to have conversations with strangers from two different races: Caucasians and Arabs, without being told the real intention of the research. The behaviours of the participants and the pedestrians were recorded for future data analysis, alongside their experiences in each conversation. After all the conversations, the participants completed an implicit association test for racial bias, a questionnaire about the their attitude toward Arabs and a questionnaire about their personality for ground truth. Finally, the participants were debriefed about the real intention of the study and then they gave the consent for recording, utilising and publishing the data.

3.2.1 Study Design

The study intends to answer 4 research questions:

a. Can we predict the perceived friendliness by participants' from their nonverbal behaviours?
b. Can we predict the participants' emotion, including level of frustration, arousal and valence from their nonverbal behaviours during conversations?
c. Can we predict the participants' personalities from their nonverbal behaviours during conversations with strangers?
d. Can we predict the participants' racial prejudice from their nonverbal behaviours facing two different races?

In total 20 participants were recruited by emails, posters, flyers and face to face recruitment. Among them, 7 were males and 13 were females. The participants were required to be native Europeans or Americans who can speak fluent German and English and between the age of 18 and 35. Additionally, only people who don't wear glasses were recruited. The capability in English and German was required because the majority of the interlocutors were German citizens, while the instruction language for the experiment was English. Furthermore, the eye tracker fits only if the person doesn't wear glasses. They could only choose from pedestrians from the Caucasian or Arabian group.

3.2.2 Procedure Design

Considering the requirements described above, a within-group method was taken and the plan of the procedure was as follows.

	Experiment Procedure (2.5 hours)			
Name	Introduction	Recording session	Ground truth acquisition	Conclusion

Description	Cover story;
	Introduction and instruction;
	Equipment Calibration;
		Converse with multiple strangers in public places;
		Questionnaire about conversation experiences;
			Implicit Association Test;
			Personality Questionnaire;
			Racial Prejudice Questionnaire;
				Debriefing;
				Consent form;
				Reward

Duration	30 mins	80 mins	30 mins	10 mins

Table 1: experiment schedule

3.2.2.1 Introduction Stage

First was the introduction stage, in which a cover story about the experiment was given. The participants were informed that the recordings were intended to build an automatic analysis system that interprets emotions from people's behaviours during conversations; in order to study the effect of cultural differences on the relationship between emotions and nonverbal behaviours, we had selected Caucasians and Arabs as two target groups; due to the fact that recruitment was difficult and that rejection or acceptance were assumed to have effects on people's emotional state, we had designed the task to be recruiting new participants for the experiments on the street while wearing a set of recording devices. Such a cover story was used so that it would not alter the intuitive behaviours of the participants.

In order to record the nonverbal behaviours of the participant and the interlocutor, each of the participants was equipped with one Intel RealSense depth camera, one Pupil Labs eye tracker and one headworn Beyerdynamics microphone. Time for adaptation and calibration of the devices were assigned, and they were given detailed instructions about the suitable environments for recording data, such as the lighting and noise conditions. In order for them to complete the tasks when strictly following the requirements, an instruction manual with a flowchart, an oral explanation and a rehearsal of the process before the real experiment were carried out. This introduction stage takes around 30 minutes.

The introduction and instruction manual can be found in appendix 1.

3.2.2.2 Recording Stage

In the recording session, the participants were required to walk around in the campus, where there were crowds of people from mixed backgrounds and had conversations with multiple strangers from the two ethnic groups. They had the full freedom in choosing which persons to talk to, as long as every single conversation lasts for around 3 minutes and they talk to a balanced number of people in terms of gender and ethnic groups. The reason for requiring them to have individual conversations and both of the people in standing pose is that in the pilot stage of the recording, these situations had led to partial occlusion of the pedestrian, and even worse, influenced the important feature in the predictive model: mutual distance. The participants were requested to ask for explicit consent from the pedestrians by reading aloud a given paragraph of notification. The consent gives permission in having the conversation while being recorded by microphone and cameras, and using the data for scientific purposes and any potential scientific publication in anonymised form. Furthermore, for those pedestrians who agree to be new participants in the study, their personal information were also documented (appendix 2).

After each conversation, the participants were requested to fill in a short questionnaire, which asked for his/her experience in the previous conversation. The questionnaire can be found in appendix 3. These questions were asked in order for per-conversation predictions, such as the perceived friendliness in the previous conversation, or arousal, valence and level of frustration of the participant during the conversation. Additionally, providing the pedestrian agrees to be a new participant in the experiment, a personal information form would be filled in in order for future contact. The details of the two forms can be found in the appendix 2 and 3. This recording stage lasts for approximately 80 minutes.

3.2.2.3 Ground Truth Acquisition and Debriefing Stage

The recording session captured the natural behaviours of participants during conversations with different races of people, as well as people's emotional states after each conversation, while the later stage measured the general attributes of people in standard psychological methods, such as level of racial prejudice and personality. In this ground truth acquisition stage, the participants were first asked a brief question: What do you think is the intention of the experiment? This question is to make sure that in the recording session, they were unaware of the real research intention, therefore we could attain the assumption that they have behaved out of their natural instinct.
Secondly, they were instructed to complete an implicit association test about their racial prejudice level toward Arabs. The test was a standard example from the Inquisit software [2]. The participants needed to select from different pairs of concepts and react as soon as possible. Their speed in pairing concepts, for instance, "Arab-negative" and "Arab-positive", revealed their opinions or attitudes toward the two racial groups. Next, an explicit racial prejudice questionnaire measuring individuals' racial prejudice toward Arabs [67] was filled. The questionnaire consisted of 42 Likert scale questions regarding people's perceptions and attitudes about Arabs. These two tests were to measure the levels of the participants' implicit and explicit racial prejudice toward Arabs, so that we could use as the ground truths. Additionally, they were requested to fill in a NEO-FFI personality test [68], where 60 questions were designed to evaluate the participants' personality traits in 5 dimensions: extraversion, agreeableness, conscientiousness, neuroticism and openness to experience. These questionnaires measured the general attitudes and personalities of the participants, which can be used for per-person prediction.

Later, the participants filled in their basic personal information, such as age, gender and subject of study, so that we could have an overview of the composition of our user group. After that, the participants were informed about the real intention of the experiment, and the experimenter debriefed them with a consent form, which included explanations of the real intention of the study and asked for permission about the usage of data for scientific purposes. This stage lasted for around 40 minutes.

Finally the participants were rewarded directly in person.

[2] Inquisit 5 [Computer software]. Retrieved from http://www.millisecond.com.

Chapter 4 Data Collection

In this section, a brief description of the devices will be first given, including the microphone, the Intel RealSense camera, the Pupil Headset, two recording laptops, and their accessories. Next, the corresponding recording software of each device will be introduced. Later comes an explanation of the data acquisition process, which can be divided into behaviour recording and ground truth annotation. Finally, a short summary of the recorded data will be given.

4.1 Apparatus

The recording set includes a Pupil Headset for eye movement and world view recording, an Intel RealSense camera for RGB and depth information of the world view, and a microphone for capturing the utterance of the participant. The set can be seen below:

4.1.1 Pupil Labs Headset

Pupil Labs is a platform for eye tracking and egocentric vision research[3]. Pupil Headsets are plug and play USB devices designed for flexible and mobile recording of the user's field of view and eye movements. The 3d printed frame can be geared with different combinations of cameras, such as one world camera and one eye camera for a monocular setting, or two eye cameras and one world camera for 3d binocular setting. Additionally, microphone can be connected to the headset so that the speech of the wearer can also be recorded.

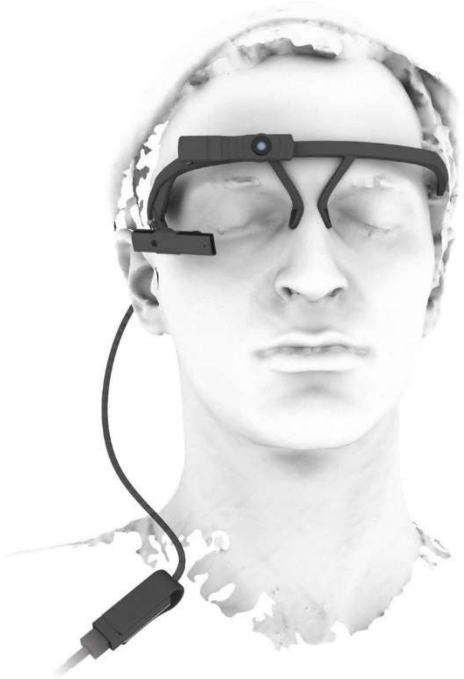


Figure 1: Pupil Labs Headset Illustration

[3] https://pupil-labs.com/pupil/

Besides the flexibility due to modularization in the hardware, the options in the open source software provides functionalities to suit diverse needs. The software Pupil Capture is for receiving, synchronizing and recording the video streams from cameras in real time, and the Pupil Player does visualization and simple analysis on the recorded data. The software is supported on Linux, MacOS and Windows platforms. The most useful documentation and forums about the product is its github page and the google forums.

4.1.2 Intel RealSense Camera

Due to the reason that there should be no restrictions on participants' body movement and activity, the recording device needs to be light and easy to carry. A few plans were thought of. A GoPro is a good option for its portability and stability, however, it only records the RGB video, without providing the depth information. Kinect is an alternative that not only gives depth information, but also does accurate and stable tracking of the body parts. However, it is only suitable for fixed location or limited movements, due to its size and need of power supply.

The long-range world-facing Intel RealSense Camera R200[4], however, captures both the RGB and depth information of the world view. It has 3 cameras providing RGB (color) and stereoscopic IR to provide depth information. Two main functionalities of R200 camera are: tracking/localization, which does real-time estimation of the camera's position and orientation using depth, RGB and IMU data; 3D volume/surface reconstruction, which represents in real-time digitally the 3D scene observed by the camera.

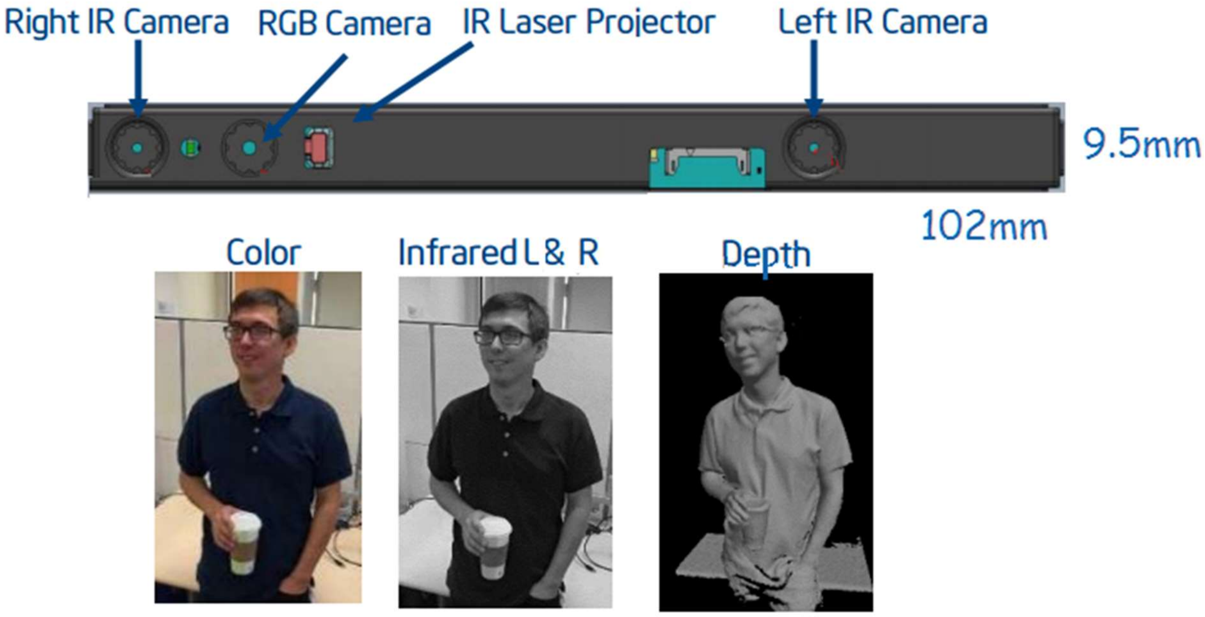


Figure 2: RealSense Camera Composition and Functions

To attach the Intel RealSense camera firmly to the participant's body and have a unoccluded frontal view, a chest mount was used to fix the camera.

[4] https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html

4.1.3 Beyerdynamic Microphone and Its Accessories

Due to the experiment design that the users will walk around freely in the university and have conversations in public areas, the device to record the user's speech needs to be easily portable, either doesn't need additional power supply or can be powered by usb hubs, and easily attaches to the body, while at the same time reserve the speech quality . A few solutions have been considered: Usb microphone directly connected with a laptop, recording with Android phone with professional recording applications, and wireless headphone which transmits signals via WiFi transmitter and so on. Considering the quality and stability of the recording, the final plan settles on a Beyerdynamic headworn condenser microphone which connects to the laptop via a condenser microphone adaptor and a SHURE XLR to USB adaptor.

4.1.4 Recording Laptop

A ThinkPad T460 laptop and a Dell XPS13 laptop were used for powering the devices and running the recording programs. The reason for using two laptops was that using one laptop turned out to be unstable, therefore the recording software crashes after some random time, ranging from 20 minutes to one hour. We went through a tedious process to find out the reason for crashing: monitoring the CPU usage and temperature of the laptop, connecting the devices to different USB controllers in the laptop, disabling the recording of the eye camera, and assigning the devices to two laptops to distribute the load.

The results revealed that when we split the devices to two laptops and utilize all USB controllers of the laptops, we were able to get the Intel RealSense depth camera running with 1080P at 30FPS recording RGB and depth information, the world camera of the Pupil Labs Headset running with 1080P at 15FPS and eye camera running with 480P at 90 FPS, together with the microphone, without the risk of halfway crash. The error message before the crush also showed that the software was not able to acquire frames from the camera after running for a while, and a continuous failure to retrieve the image had led to the crush of the software. Additionally, when we ran all the three devices on the same laptop, the real frame rate of the cameras dropped to less than half of that in the setting, for instance from 30FPS to 13 FPS, and fluctuated heavily. Therefore we have the speculation that the reason for such a crush is that the size of data flow exceeds the bandwidth limitations of the laptop, therefore causing resource competition between the devices and hence continuous failure in one channel.

4.2 Software

In the following subsections, the software used for the acquisition of recording in different channels will be introduced.

4.2.1 Pupil Capture and Player

Pupil Capture reads the video streams from the world camera and the eye camera of the Pupil Labs Headset. It detects user's pupil position, tracks user's gaze, detects and tracks markers in the environment, records video and events, and streams data in realtime. Several different calibration methods including screen marker calibration, manual marker calibration and so on are provided by the software. Similar to other video recording softwares, it also support different frame rate and resolution for the streaming. Furthermore, there are plugins which enable additional functionalities, such as synchronization of multiple input sources, streaming data over the network, blink detection and so forth. The output of the software consists of the recorded videos, the timestamps of the images shots, the pupil data, the detected blinks, the calibration data and general information about the clip.

Pupil Player is the software to playback the recorded video. It is a media and data visualizer. It also supports fundamental processing of the recorded data. Exporting from Pupil Player, we can get excel files which include pupil position recognition, gaze position estimation, fixation detection and so on.

The important and relevant features of the software are as follows: streaming, synchronizing and recording videos from both the world camera and eye camera, detecting pupil position of the eye, detecting fixations and blinks in eye movements and estimate the gaze position. Therefore, the important features of eye activity can be either directly retrieved or calculated from the results of Pupil Capture and Pupil Player, including fixations, saccades and blinks.

One major drawback of Pupil Capture is that the timestamps returned by the software is not the world timestamp of the recording laptop. Instead, it adopts an arbitrary start point, which makes it hard to synchronize the eye tracking data with other channels of signals. The same applies to the videos recorded with RealSense. One option is to call a specific function at the beginning of the recording. However, calling with command line does not facilitate the calibration stage before the real recording. Hence, manual annotation was used to synchronize the videos recorded with Pupil Headset and RealSense.

4.2.2 Script for using RealSense

To use the RealSense camera, the correct versions of Intel RealSense Depth Camera Manager and Intel RealSense SDK need to be installed. The DCM is intended to exhibit interfaces of streaming video from the camera for in color and depth view. The SDK is capable of a set of functionalities in data processing, such as facial recognition, hand gesture recognition, background removal, 3D scanning and so on.

However, the default functionality of the DCM only enables real-time camera exploration, or playback of the recorded file, while we would prefer a live calculation of mutual distance. Therefore, to stream the color and depth video and record them, we need to call the functions in programs. We have chosen C++ as the programming language and the development environment was Visual Studio 2015.

The data format of recorded videos from RealSense SDK is .rssdk, which includes the RGB and depth information of the scene. However, for this device, the expected function is to measure the distance between the participant and the other speaker. As a result, besides saving the video, the script was programmed to keep track of the system time, the location of the detected face in the view and the distance of the detected face from the camera. The output of this program is a text file of the above information.

4.2.3 Audacity

Audacity[5] is an open source, cross-platform audio software for multi-track recording and editing. The important relevant features of Audacity are: recording from microphone, line input, and USB/Firewire devices; record computer playback; create and export .mp3 or .wav files; supports diverse sound quality; capable of editing on effects such as noise reduction, high pass and low pass filters, notch filter and so on; basic speech analysis such as viewing and plotting spectrum and contrast analysis.

4.2.4 Inquisit

Inquisit [6] is a general purpose psychological experimentation application for designing and administering psychological experiments and measures. It can run a given script locally on a Windows PC or Mac, or it can host online experiments over the web. The software can be utilized to implement a wide range of experiments, such as reaction time tasks, psychophysiological experiments, attitude measure, surveys and so on. A considerable set of experiments were already programmed and provided to users, including the relevant task: Arab-Muslim IAT [69], which measures the implicit racial prejudice toward Arabs.

4.3 Data Acquisition

Since one of the aims of the experiment was to predict people's level of racial prejudice toward Arabs, we have reduced our participant group to German speaking Caucasians, in order to reduce the complexity. The Pupil Headset requires that the user should not wear glasses, therefore the prerequisite for the participant turned out to be strict: 18-35 years old, a Caucasian who speaks fluent German and English, and doesn't wear glasses. As a result, the participants were recruited in many ways and cost much efforts: poster, flyers, in-person recruitment and group emails.

To acquire the nonverbal behaviour data, an experiment has been designed. Essentially, the experiment is intended to record the natural nonverbal behaviors of the participant when confronting and conversing with strangers from different races, while not letting the participant know the real intention of the study. Therefore, despite the need for a complete plan for multimodal signal acquisition, we also need to come up with a cover study to prevent the participant from noticing that the experiment in fact studies his or her behaviors toward different races.
The following picture is a demonstration of a user wearing the full gear set.

[5] Audacity(R) software is copyright (c) 1999-2016 Audacity Team. [Web site: http://audacityteam.org/. It is free software distributed under the terms of the GNU General Public License.] The name Audacity(R) is a registered trademark of Dominic Mazzoni."

[6] Inquisit 5 [Computer software]. Retrieved from http://www.millisecond.com.

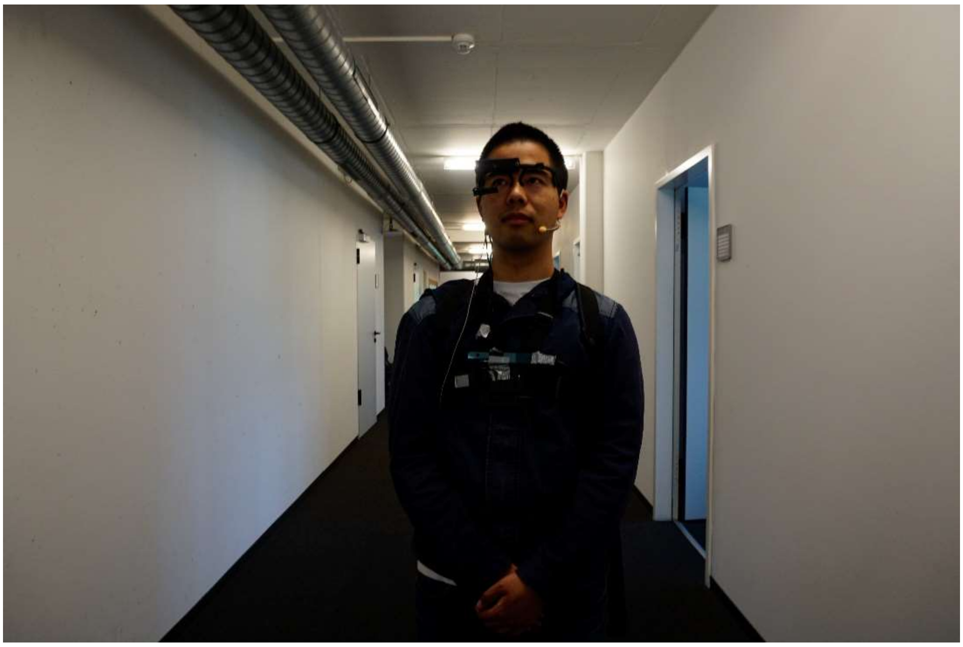


Figure 3: illustration of a user wearing the full gear set

Furthermore, the multimodal recording framework only captures the nonverbal behaviours of the participants, but not the ground truth. To access the fact about the participant's racial bias and personality, or his/her current emotional state, or the perceived friendliness of the pedestrians, a few psychological methods were used. For the topic of racial prejudice, the implicit association test was implemented to evaluate the participant's level of implicit racial prejudice, while the New Anti-Arab Attitudes Scales (appendix 4) measured the explicit racial prejudice toward Arabs; for the topic of personality, a NEO-FFI questionnaire (not included in appendix due to possible intellectual property right issues) for personality were completed by the participant. Besides the above mentioned general measures at the end of the experiment, other measures were taken after each conversation: the participant was required to fill in a form about his/her experience during the previous conversation, including perceived friendliness of the pedestrian, his/her current arousal and valence level, how comfortable he/she felt during the conversation, and so forth.

After the participant came back from the recruitment, he/she would be first shortly interviewed for his/her experience in the conversations. The primary reason for such a short interview is that we needed to check whether the participant had already noticed or guessed about the real intention of the experiment. An additional reason was to use this interview for enhancement of the experiment design.

Next, the participant was required to complete the implicit association test of racial prejudice toward Arabs on a laptop individually in a closed room, so that he/she was not influenced by any environment factors. The participant was requested explicitly that he/she should read the instructions carefully and do the test out of instinct.

After the participant finished the test, two more questionnaires were offered. The first one was the NEO-FFI personality test and the second one was the New Anti-Arab attitudes scale. The participant was required to be seated individually in a room and fill in the questionnaires in paper version anonymously and truthfully.
Next, the experimenter came back to the participant with a brief explanation of the real intention of the experiment. The participant was informed that he/she was told a partially true

story: besides the automatic analysis of emotional states from behaviours, we also intended to study how nonverbal behaviours reflect people's racial bias levels, or how nonverbal behaviours influenced people's perceived friendliness or emotional experience in conversations. The participant was presented with a debriefing form which gave detailed explanations about the experiment and the rights and responsibilities of the participant.

Finally, the participant received the payment from the experimenter, and was reminded again that since this was an ongoing study, he/she should never reveal the true story of the experiment to any other people.

4.4 Overview of Data Set

In total 20 participants were recruited to do the experiment, several samples were removed as a result of software failure, missing data or that the participant didn't follow the instructions strictly, depending on the features selected in the model. 7 participants were male while 13 were female. The age of the participants ranges from 18 to 30 (mean=22.95). The subjects of their study covered a wide range of faculties.

The dataset for each participant includes a video of the world view and a video of the eye view from the Pupil Labs Headset, a video of the world view from the RealSense depth camera, and an audio file from the microphone. Furthermore, 2 psychological questionnaires, a number of questionnaires about conversation experience and the results from the implicit association test are also included. The total size of all the files of each participant reaches around 100 gigabytes, depending on the duration of individual recordings.

Chapter 5 Data Analysis

5.1 Data Set Preprocessing

The known variables in the experiment are the race of the interlocutors and the attributes of the participant or the experience of the conversation. What the regressor needs to do is to predict those characteristics of participants or conversations from the social signals in the conversations. The attributes of the participant includes the level of racial prejudice and the personality of the participant. The experience of the conversation includes the perceived friendliness of the pedestrian, and the corresponding emotional states: arousal, valence and level of frustration after each conversation.

The corresponding social signals are the nonverbal behaviours of the participant or the pedestrian: the behaviours of the participant are closely related to the attributes of the participant, while the behaviours of the pedestrians influence the participant's impression of friendliness and the emotional states of the participant. We measure the social signals of both people through various channels: from the participant's eye movement, interpersonal distance, and speech data, or from the eye movements, and the facial expression of the pedestrians.

An illustration of the logic can be seen below.

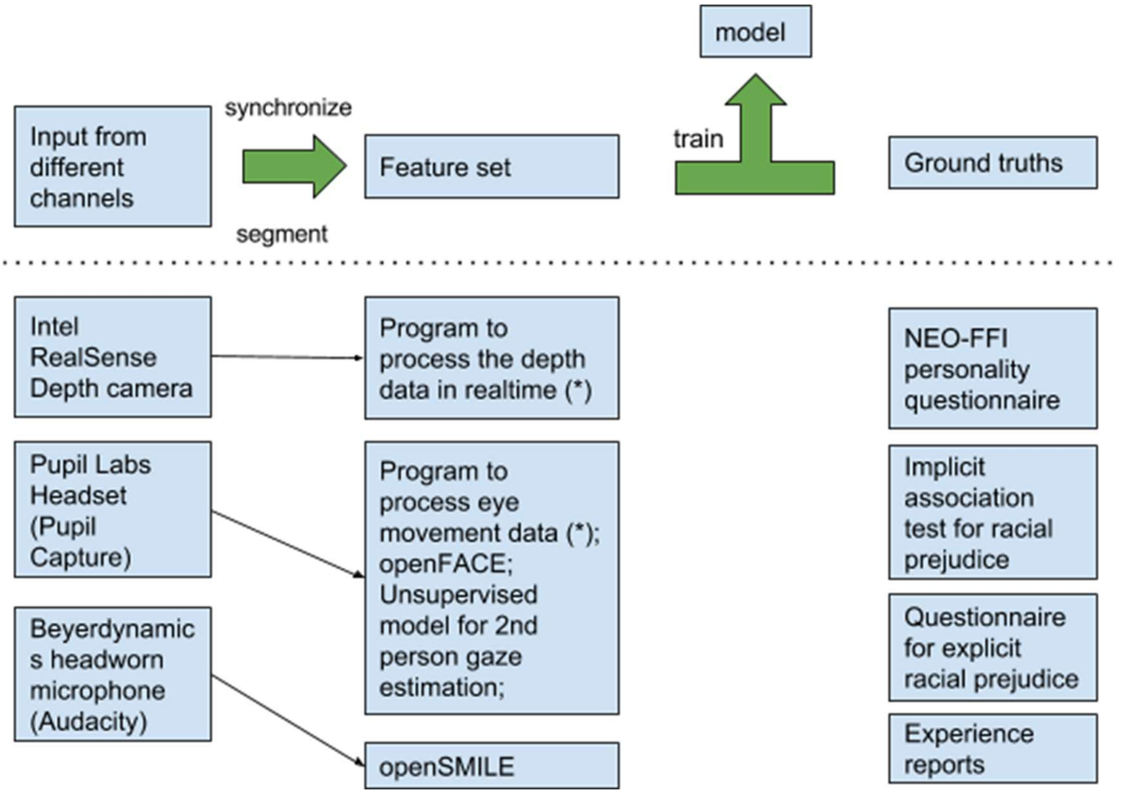

model
Input from different channels
synchronize
segment
Feature set
train
Ground truths
Intel RealSense Depth camera
Program to process the depth data in realtime (*)
NEO-FFI personality questionnaire
Pupil Labs Headset (Pupil Capture)
Program to process eye movement data (*); openFACE; Unsupervised model for 2nd person gaze estimation;
Implicit association test for racial prejudice
Questionnaire for explicit racial prejudice
Beyerdynamics headworn microphone (Audacity)
openSMILE
Experience reports

Figure 4: From raw data to features

Since the conversations were recorded in 3 different channels, the synchronization of them is significant. This was realized by a sudden and loud clap in front of the participant before the real recording. The videos from the eye camera and world camera from Pupil Lab Headset were automatically synchronized. For videos from the world camera from Pupil Labs headset and the RealSense camera, the timestamp of clapping was marked, since with one frame difference, the openness of the palms had changed. For the audio, the clapping proved to be a sudden burst of sound in the sound track, and we marked the beginning of this sound to be the exact point of clapping.

Each recording of an experiment normally consists of 6 to 10 conversations, with the same participant but different pedestrians. The first step would be to cut conversations out of the whole recording. Based on observations from the video, the start of each conversation is defined to be the time when the pedestrian agrees, either verbally or nods, to spend a few minutes in the conversation; the end of each conversation is defined to be the time before the pedestrian makes any movements to walk away, or before the pedestrian starts to fill in their personal information. The audio was therefore cut according to the corresponding timestamps in videos.

Then for each conversation, a set of features were calculated in windows.

5.2 Processing of psychological Measurements

In this section, a general description of all the psychological measurements utilised in this experiment will be given. In order to prevent wrong understanding of statements in the questionnaires, the New Anti-Arab Attitude Scales, NEO-FFI personality questionnaire, and the implicit association test toward Arabs are either chosen as or translated into German version, which is the native language of the majority of the participants.

5.2.1 Explicit Racial Prejudice

The New Anti-Arab Attitude Scales (appendix 4) is a measure for evaluating individual's level of explicit racial prejudice toward Arabs, which was proved to have satisfactory psychometric properties and shared evident correlation with the adapted Modern Racism Scale [70]. The measurement is designed to adapt to the anti-Arab prejudice in the European context.

The questionnaire consists of 42 statements, and the participants were required to tick their extent of agreement to each statement, ranging from 1 (strongly disagree) to 7 (strongly agree). In order for the participants to understand the questionnaires thoroughly, the questionnaire was translated into German by a native German speaker.

To group all results to a single value, the solution was to add up the scales in each statement and reverse the negative loadings.

5.2.2 Implicit Racial Prejudice

The implicit association test about racial prejudice toward Arabs was carried out with Inquisit 5. The participants were required to press keys on a laptop in order to select the concepts they deemed as paired as soon as possible. In this experiment, common names including Caucasian names or Arabian names appear on the screen, and the user needed to pair the name with positive or negative adjectives as instructed. The reaction time differences among different pairs of combinations were calculated, for instance the time difference between pairing a Caucasian name with happy and pairing an Arabian name with happy. However, the pilot study shows that some users were unable to tell which name belongs to which group. Therefore, a minor change has been made in the code to replace the tricky names with typical Arabian and Caucasian names.

Under a series of manipulation, such differences led to a categorical value among high, moderate, and low racial prejudice toward Arabs/Muslims, as well as numeric value representing the level of racial prejudice.

5.2.3 Personality

The NEO Five-Factor Inventory (NEO-FFI) is a measurements of individual's personality from five basic personality perspectives: extraversion, agreeableness, conscientiousness, neuroticism and openness to experience. Neuroticism is defined to measure individuals' tendency to be moody and experience negative feelings such as anxiety, worry, fear, frustration, envy, anger, loneliness and so on. Extraversion depicts how outgoing, talkative and energetic an individual is. High agreeableness characterizes those who demonstrate behaviours that are perceived as kind, sympathetic, cooperative, warm and considerate. Conscientiousness is a personality trait of being careful and vigilant. Lastly, openness to experience means a person's appreciation for art, emotion, adventure and a variety of experience [7].

In total 60 items cooperate to infer the individual's personality in the above mentioned 5 dimensions. The items use a format of Likert scale, ranging from 1 (strongly disagree) to 5 (strongly agree). Single values for each dimension were therefore calculated from only the related items.

5.2.4 Emotional State

The emotional state of participants were measured within the two dimensional valence-arousal model [71]: valence, which means pleasantness value, and arousal, which means bodily activation. The range of scores is between 1 and 9. A plot vividly depicting the different levels of arousal and valence values was taken for measurement, as can be seen in appendix 3.

Additionally, the frustration level of the participant was measured in a range of 1 (not frustrated at all) to 5 (very frustrated).

5.2.5 Perceived Friendliness

The participant was also required to rate his/her perception of the friendliness of the pedestrian during the conversation. The scores are from 1 (not friendly at all) to 5 (very friendly).

[7] https://en.wikipedia.org/wiki/Revised_NEO_Personality_Inventory

5.3 Feature Extraction

The feature set was concluded from summaries of related work, intuitive discussions about relevant affective signals during conversations, and notes from manual annotations of conversations (see appendix 9). 3 annotators were recruited to annotate a same subset of videos from the dataset and rate their perceived friendliness of the pedestrian. The clues for giving the scores provided by the annotators are summarised in appendix 6.

In this section, a detailed description about the feature set will be given. Please mind that the features from all channels were calculated on a shifting window basis. Specifically, the conversations are divided into windows of 10 seconds each, and the remaining segment smaller than 10 seconds will be abandoned.

5.3.1 Audio

5.3.1.1 Software – openSMILE

The openSMILE software [72] is the Munich open-Source Media Interpretation by Large feature-space Extraction (openSMILE) toolkit. It is a modular and flexible feature extractor for speech signal processing and machine learning applications. Despite the functionalities in feature extractions from audio signals, openSMILE has the following relevant advantages which are asserted to be rare in other similar software: i) it supports batch processing for large data-sets and extract features incrementally; ii) openSMILE provides access and recording and visualisation of intermediate data during the processing.

5.3.1.2 Data Processing

openSMILE provides different types of configuration files which serve to directly extract basic features from audio files. The features can also be calculated with shifting windows. In our study, the extended version of the Geneva Minimalistic Acoustic Parameter Set (eGeMAPs) [73] has been adopted. The feature set was established especially for affective computing and includes 18 low-level descriptors (LLDs) in several groups of parameters: frequency related parameters (F0, formants frequencies, etc.,), energy/amplitude related features (shimmer, loudness, harmonics-to-noise ratio, etc.,), spectral (balance) parameters (alpha ratio, Hammarberg Index, spectral slope, formant relative energy etc.,), and temporal features (rate of loudness peaks, mean length and deviation of voiced regions, etc.,). Combining with supplementary arithmetic calculations on the LLDs and the temporal features such as the rate of loudness peaks or the mean length of voiced regions, the whole minimalistic contains 62 parameters in total. The extended version of the parameter set introduces cepstral parameters and other dynamic features, making the feature size to reach 88.

To extract the above mentioned eGeMAPs feature set from speech audio, the parameters in the configuration file for execution has been modified, especially the size of the window and the shift for computing the features. In order to read the output file which is in arff format, functions have been written to feedforward the features. The detailed explanation of the feature set can be found in the work of Eyben et al., [74].

5.3.2 3d Depth Camera

The returned data from the recording program includes the machine timestamp of the current frame, the locations of detected faces in the view and their distances from the camera.

The features were calculated for each conversation, including: the average of mutual distances, the variance of distances, and the orientation the participant takes toward the pedestrian. The first two features determined by taking the arithmetic mean and standard deviation of mutual distance across time, while the last feature divide the distance of the pedestrian's face from the scene center by the current mutual distance. Intuitively, the distances were intended to show the intimacy of the two speakers, while the orientation reflects the participant's friendliness or openness to the pedestrian.

5.3.3 Synchronized World and Eye Camera

The output of the Pupil Capture software include the raw video of the eye camera and world camera, raw data consisting of detected pupil positions, timestamps of each frame from the world and eye cameras, as well as the basic information such as the duration of videos, the start and end time of the recording, and so on.

Pupil Player was used to read the above files and yield excel files which consists of the positions of pupil, the positions of gaze and start and end of fixations.

5.3.3.1 Eye Movement Features

The processing of the fixation data and pupil or gaze positions was based on the work by [75]. A program has been written to produce the features such as the average, variation, maximum or minimum of the pupil diameter, and that of the duration, frequency and amplitude of fixations and saccades.

5.3.3.2 Higher Level Features

The following synthetic features were computed by extensive manipulation and combination of the original gaze and world view data.

- Facial Expressions are deemed to indicate the friendliness of a person, hence influencing the experience of conversations. As a result, the facial expressions of pedestrians were extracted with the help of the open source facial behaviour analysis toolkit: openFACE. This tool is not only able to do facial landmark detection, head pose estimation and facial action unit recognition, but also eye-gaze estimation. Here we have utilised the facial action unit under the facial action coding system [76] to reconstruct the emotions hidden in facial expressions. Specifically, the selected facial action units are AU1 (inner brow raiser), AU4 (brow lowerer), AU6 (cheek raiser), AU12 (lip corner puller), AU 14 (dimpler) and their combinations such as smiles. These units are considered to be prominent in positive or negative emotions.

Further features have been computed from the eye movement features.

- The location of the gaze in the scene assist to reveal whether the participant was looking at the pedestrian's face region, or specifically the eye region, nose region or mouth region. This is realised by reading the facial landmark detection results from openFACE and defining bounding boxes for each facial regions of the pedestrian. Next, we matched the current gaze location in the scene with the detected face region of the pedestrian. Expanded bounding boxes were drawn to cover the errors in face region recognition due to the small region and much noise from dynamics.
- The duration and dynamics of the mutual gaze between the participant and the pedestrian were also taken into account. This is achieved with a unsupervised model [77, 78] for estimating gaze location from a second person perspective, therefore enabling the estimation of the pedestrian's gaze location. A program has been written to match the two gaze locations and calculate the features for mutual gaze.
- Length of the conversation is highly correlated with the participant and pedestrian's willingness to continue conversations, therefore relates to the experience of the conversation.
- The gender of the pedestrian is also taken into the feature set, considering gender effect.

A detailed table of the higher level features can be seen below.

Name	Description
p1	Percentage of time that the participant's gaze is on the pedestrian's face
Avg_p1_duration	Mean duration of the participant's gaze on the pedestrian's face
p2	Percentage of time that the pedestrian's gaze is on the participant's face
Avg_p2_duration	Mean duration of the pedestrian's gaze on the participant's face
percent1	Percentage of mutual gaze when participant's gaze on the pedestrian's face
percent2	Percentage of mutual gaze when pedestrian's gaze on the participant's face
Avg_mutual_duration	Mean duration of eye contact
shift_x	Orientation of the participant, i.e. distance of pedestrian's detected face from the center of the scene divided by mutual distance
AU6	Average intensity of cheek raise across time
AU12	Average intensity of pulling lip corner across time

smile	Average intensity of AU6*AU12 across time
smile_peaks	Number of peaks found in smile divided by conversation length
peak_mean	Mean intensity of peaks found in smile across time
frown	Average of presence or absence of AU9 (nose wrinkler) across time
AU1	Average intensity of inner brow raise across time
AU4	Average intensity of lowering brow across time
AU14	Average intensity of dimple across time
ptg_switch	Number of switch between on-face and off-face gaze divided by conversation length
ptg_eye	Percentage of time that participant looking at eye region of the pedestrian
ptg_nose	Percentage of time that participant looking at nose region of the pedestrian
ptg_mouth	Percentage of time that participant looking at mouth region of the pedestrian
distance_mean	Mean mutual distance across time
distance_var	Variance of mutual distance across time
distance_slope	The regression coefficient of the distance versus time
gender	Gender of the pedestrian
length	Length of the conversation segment

Table 2: higher level feature set

The full feature set include the above explained higher level features (26), the eye movements features (33), as well as the speech features as decoded in eGeMAPs (88) norm.

5.4 Pipeline of Model

In this section, several graphs will be shown to clarify the steps that had been taken to preprocess the feature set and train the model.

5.4.1 Preprocess Chart

First is the preprocess chart, which starts from the feature sets from multiple input sources and the target value set, and ends with a pandas dataframe consisting of the full feature set and target values.

As can be seen below, after merging into a full feature set where the unit is a window, samples were filtered based on the approach to take. If we use each conversation as a sample, then the features are average among windows for each participant; if we use segments as samples, only one window from each participant will be chosen as the new samples. Later the outliers will be removed, but null values will be kept, followed by the interpolation of the features person-wise. Finally, the outliers and null values will be removed, since now a null value means there is no valid features of the other recordings of the participant, therefore making it useless.

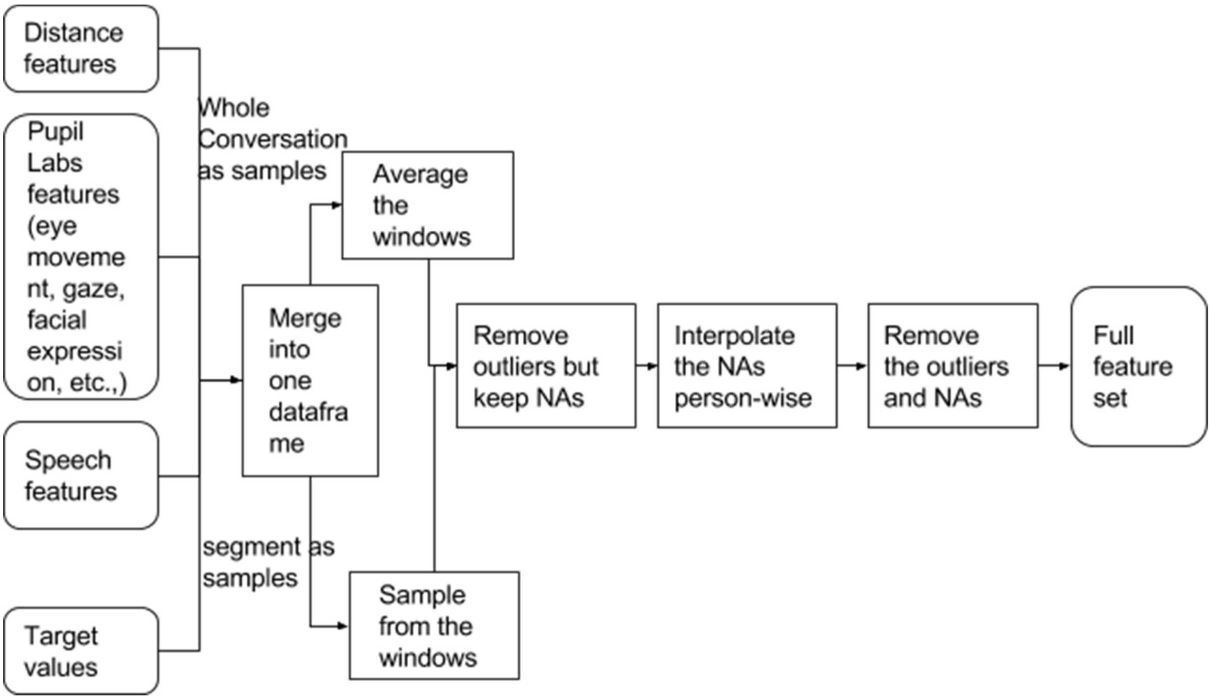

Distance features
Pupil Labs features (eye moveme nt, gaze, facial expressi on, etc.,)
Speech features
Target values
Whole Conversation as samples
segment as samples
Merge into one datafra me
Average the windows
Sample from the windows
Remove outliers but keep NAs
Interpolate the NAs person-wise
Remove the outliers and NAs
Full feature set

Figure 5: Preprocess the feature set

5.4.2 Reasoning of the Complexity

Second is the illustration of the model training. Due to the dimensions of the options along the pipeline, the training process could be very complicated:

a. Using the original ratings or the third person annotations

In the early stage of the model training, we have found out that the models seemed unable to fit the dataset we have. A speculation has emerged that the ratings from the participants were too random to predict. Consequently, two small studies were designed and carried out so as to test the reliability of the scores given by the participants, as well as generate a more consistent series of ratings for the conversations.

In the first experiment, 3 annotators were recruited to rate their impression of the friendliness level of different video recordings of 13 conversations. The conversations were from 4 different participants, and the annotators were uninformed of the ratings given by the participants. As it happens, the average correlation coefficient of the ratings from the 3 annotators was moderate (mean= 0.46), while the correlation coefficient of the ratings between the annotators and the participants were low (mean=0.24). The results to some extent justifies that the prediction of ratings from participants involves much complication. The details of this study can be found in appendix 9.

In the second experiment, an annotator was recruited to rate all the video recordings of conversations, without knowing the scores given by the participants. Such a study aims to generate a consistent rating of all the conversations.

As a result, for the task of predicting friendliness, an extra series of target values need to be fitted and evaluated.

b. Per-conversation or per-person prediction

In line with the norms of predictive models in similar cases, scores were predicted in three fashions: user-independent, user-specific, and user-adaptive. User-independent denotes the approach that the training data and the test data are from different data groups. This corresponds to the situation that the system should be able to predict for a completely new user. User-specific denotes that approach that the training data and the test data are from the same data group, meaning that the system gives prediction for a recording based on only the previous recordings of the same person. User-adaptive is a combination of the two approaches, where the system gives predictions on the basis of the previous recordings of the current person as well as the recordings from other users.

It's logical to conclude that the user-independent approach can be used to carry out both per-conversation and per-person predictions, while the user-specific and user-adaptive approach only works for per-conversation tasks.

Therefore, for different types of tasks, the conditions to compare are different.

c. Comparison of different feature sets

As has been explained before, the feature sets can be divided into several groups. Therefore there were a few options and their combinations to compare the contributions of different feature sets, as can be seen below:

- Features from eye camera
- Features from world camera (including depth)
- Features from microphone
- Eye camera together with world camera
- Full feature set

d. Comparison of different algorithms and their hyperparamter tuning

A number of algorithms were implemented to produce the regression model.

Support vector machines [8] are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis. The penalty parameter C of the error term and the margin tolerance value of distance epsilon were set between 0.001 and 1000 on a logarithmic scale.

Ridge regression [9] is a type of linear regression that imposes penalty terms on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares, which leads to the shrinkage of coefficients in the linear model. Furthermore, adding polynomials combinations of the features makes the regression of nonlinear relationship with the linear models possible. The regularization strength term α was set to range between 0.001 and 1000, and for polynomial ridge regression, the degree of polynomial was set to 2. The formula for the penalized residual sum of squares can be seen below.

$$\underset{w}{min} ||Xw - y||_2{}^2 + \alpha||w||_2{}^2$$

Decision trees (DTs) [10] are a non-parametric supervised learning method for classification and regression. The model predicts the value of the target variable by learning simple decision rules inferred from the data features. The maximum depth of the tree to be considered ranges from 2 to 10, and the maximum number of features considered during split was set to the ‘sqrt’ mode, where the value was set to the square root of the number of features.

Random Forests (RFs) [11] is an ensemble learning method for classification and regression. It constructs multiple decision trees during training time and outputs the mode of the classes or the mean of the prediction of the different trees, therefore correcting the disadvantageous overfitting habit of decision trees. The maximum depth and the maximum number of features of the trees were set to the same as that adopted in decision tree, while the number of estimators was set to 10.

Nearest neighbours (KNN) [12] is a method that finds a predefined number of training samples closest in distance to the new point, and predict the label from the interpolation value of those points. The number of neighbours was tested between 3 and 8, and the weight function used in prediction was tested between uniform or by the inverse of their distance.

The gradient boosting machines (GBM) [13] is an ensemble of multiple weak prediction models, which in most cases are decision trees. Unlike random forests, it builds the model based on the previous model, by continuously adding shallow trees into the sequence. The

[8] https://en.wikipedia.org/wiki/Support_vector_machine

[9] https://en.wikipedia.org/wiki/Tikhonov_regularization

[10] https://en.wikipedia.org/wiki/Decision_tree

[11] https://en.wikipedia.org/wiki/Random_forest

[12] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

[13] https://en.wikipedia.org/wiki/Gradient_boosting

number of estimators to be trained was set to 100, 200 and 500, while the maximum depth for each tee were tested among 3, 4, 5 and 6.

Multilayer perceptron (MLP) [14] is a class of feedforward artificial neural networks. It consists of at least 3 layers and the activation function is nonlinear. The hyperparameter were as follows: the number of hidden units in each layer was set to 10, and the number of hidden layers ranged from 1 to 3; the activation function for the hidden layer had three options, the logistic sigmoid function, the hyperbolic tan function, and the rectified linear unit function.

e. Comparing whole conversations and segments of conversations

As was illustrated in the preprocessing stage, options can be made between using the whole conversation, the first 10 seconds of each conversation, or using roughly the last 10 seconds of the conversation as samples.

Therefore, to tune for the best parameters for each algorithm and compare the best performance of various algorithms could be demanding.

5.4.3 Cross validation

Here we would like to explain how the parameters were tuned and how the models are evaluated. Using Leave-one-subject-out fashion for user-independent models as an example, suppose there are 20 participants in the whole data set, below is how we split the data set.

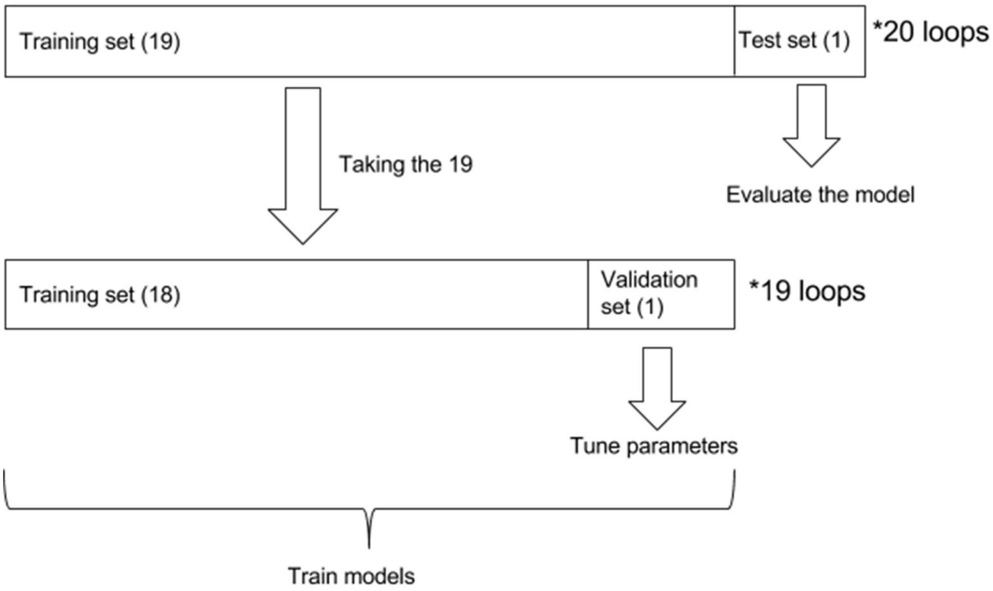


Figure 6: Illustration for Cross-validation

Each of the 20 participants will be left out for prediction, and with the other 19 participants, models were trained with a second cross-validation, where the parameters which achieves best mean performance with the validation sets will be chosen as the parameter for the outer

[14] https://en.wikipedia.org/wiki/Multilayer_perceptron

loop prediction. Taking the average of the performance from 20 different models on 20 different participants, we can have a reliable estimate of the test error of the current model.

5.4.4 Mind map for model training

It’s undeniable that if we experiment with every possibility explained above, it will lead to extremely massive work. Therefore, we have to and have designed a mind map which intends to simplify the pipeline while at the same time reserve the parts we would like to study.

First of all, regarding the whole model training as a sequential action, we have confined the flexibility from the starting point: we would start with the prediction of friendliness with the full feature set in a user-independent system, and use whole conversations as samples.

We would like to determine which algorithm is superior, and which target value produces better performances as early as possible, but in a solid way. Therefore, we will first experiment with these two options in parallel. After hyperparameter tuning within cross validation, positively, the superiority of either the original target value or the third person annotation should be obvious. The next step to do is to vary the size of conversations in the training set and plot the performance versus the training set size. This only needs to be done with two or three algorithms which produce acceptable results in the previous step. In such way, with much performance outputs which depicts the absolute errors as well as the trend of the errors, we could select the best performing algorithm for per-conversation models in a solid way.

The procedure can be seen below.

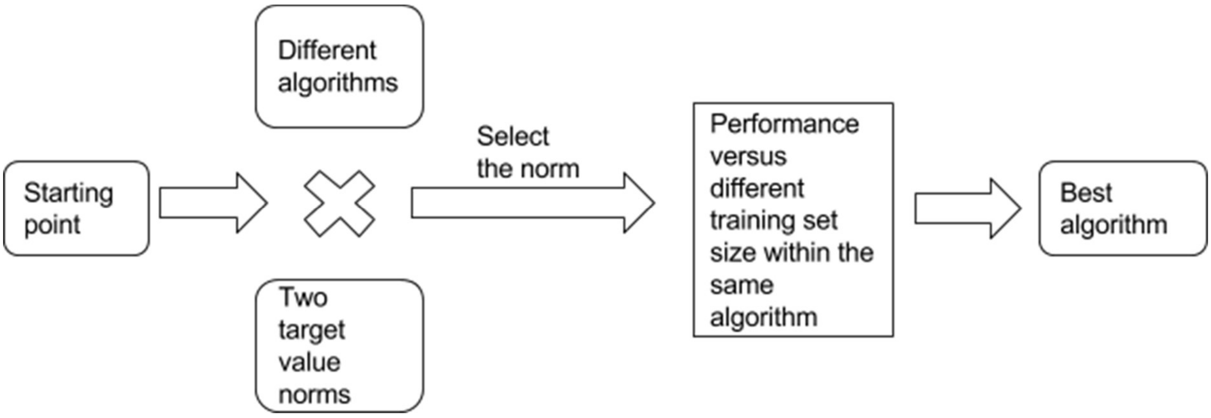


Figure 7: Simplified pipeline 1

After the above manipulations, we have decided on the algorithms and the target values to build on. Next, we could continue experimenting with the user-independent, user-specific and user-adaptive models, and pick the setting which outputs the best result. It’s worth noting that the user-adaptive model is in principle a leave-one-conversation approach, where the function train on all the other data expect for one point, and such training happens for every data point in the data set. When another cross validation is also used to tune the hyperparameters, the number of training to run could become enormous.

Later, we could continue with comparing the whole conversation or a fraction of conversation as samples, finally followed by feature selection.

The graph below starts from the ending point of figure 7.

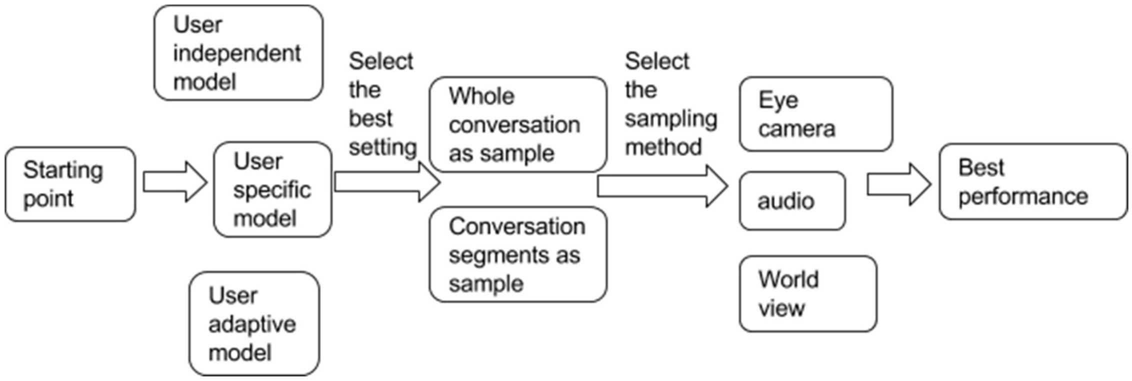


Figure 8: Simplified pipeline 2

Based on the logic given above, the results section will also follow such a mindset.

Chapter 6 Results and Analysis

In this section, the top performances of each prediction tasks will be given. Moreover, the mind map of how to select different options throughout the pipeline and how the algorithms were improved will be given in detail. To take an overview of the best performances, please jump to section 6.4.

6.1 Predicting Friendliness

In line with the mind map, a user-independent model was first trained to predict the friendliness of pedestrians both as perceived by the participants and as labelled by the annotators with the full feature set.

6.1.1 Selecting Target Value Norms and Algorithms

A dummy regressor was applied to give a baseline performance. The regressor always gives the mean value from the training data as the prediction. All the algorithms were implemented within the scikit-learn [15] library in python.

6.1.1.1 Comparing results from combinations of norms and algorithms

	Target values by participants		Target values by annotators	
Algorithm	MSE	R_squared	MSE	R_squared
Baseline	0.724	-0.52	0.802	-0.32
SVMs	**0.704**	**-0.32**	0.834	-0.19
Ridge	0.803	-0.28	**0.779**	**0.15**
Ridge+polynomials	1.103	-0.10	1.762	- 0.00005
Decision Trees	0.931	-0.12	0.817	0.16
Random Forests	0.797	-0.13	**0.770**	**0.17**
KNN	0.980	-0.16	0.872	0.03
GBM	0.922	-0.06	**0.803**	**0.20**
Neural Networks	4.108	-0.165	3.898	0.009

Table 3: Mean performances across algorithms and options for target values

[15] http://scikit-learn.org/stable/

One-by-one analysis will be done on the above performances. First of all, for models based on the target values given by the participants, only one support vector regressor seemingly achieves the same results as the dummy regressor. However, the MSE of the two regressors are extremely close to each other, while the correlation coefficient between the prediction and the truths was -0.32. Therefore, it's obvious that with the participants' target values, the system failed to train a successful model.
Now taking a look at the regression models with the annotated target values. A few algorithms generate smaller or similar mean squared errors than the dummy regressor: ridge regression, random forests and gradient boosting. Therefore, in the following sections, we will take these three algorithms as the default training model for annotated target values.

6.1.1.2 Testing with the "by participant" prediction

We would like to test whether or not the prediction with participants' target values is possible. A few things can be done: analyzing the correlations between the feature set and the target value; and seeing if the performance improves as the size of the training set increases.

Firstly, the correlation coefficients between the feature set and the participants' target values have been computed. The outcome turns out that when an absolute value of coefficient larger than 0.2 and the p value smaller than 0.05 are required, there is only one feature up to the requirements: the mean interpersonal distance between the two speakers. In other words, there is no linear relationship found among all the other 146 features. However, we cannot deny that such a phenomenon only reject the possibility of linear relationships between the feature set and the target value, while not applicable for nonlinear relationships. Therefore, we continue the training with the only promising algorithm – support vector machine.

We implemented user-adaptive models with support vector machines, which in theory should give the best performance among user-independent, user-specific and user-adaptive systems. The result is turns out that the mean squared error with the support vector machine model is 1.46, while the dummy regressor achieves MSE of 1.05. The correlation between the prediction and truth values is -0.03 (t-statistic=1.330, pvalue=0.1845). In other words, no models could be trained which outperforms the dummy regressors.

As a result, we have dropped the plan of predicting friendliness based on the scores given by the participants.

6.1.2 Selecting Type of Systems

As has been discussed in the previous subsection, 3 algorithms seemed to give promising results when predicting the friendliness value from the annotator. Here we take a further step into the different types of models, in other words, user-independent models, user-specific models and user-adaptive models.

To make a solid comparison between the two algorithms, we compared the best performance each algorithm could achieve within the 3 models. The summarizing table is as below:

	User-independent model		User-specific model		User-adaptive model	
Algorithm	MSE	R_squared	MSE	R_squared	MSE	R_squared
Baseline	0.802	-0.32	1.189	-0.44	1.233	-0.09
Random forests	**0.770**	**0.17**	**1.264**	**-0.22**	0.730	0.32
GBM	0.803	0.20	1. 666	-0.24	**0.696**	**0.40**
Ridge Regression	0.779	0.15	2.065	-0.17	0.764	0.35

Table 4: Mean performance across model types with two algorithms

It can be seen that for user-independent models, ridge regression, gradient boosting machines and random forests achieved equal or above performance as the baseline, and the random forests achieves the best performance among them. However, it is also true that for user-specific models, all of the algorithms behaved worse than the baseline. Such performances could be logical, since the training data in user-specific models are only the conversations from the current participant, which could vary between 2 and 13, therefore introducing much variance into the performance. For the user-adaptive models, all three algorithms achieve better results than the dummy regressor. Therefore, we can conclude from data that user-adaptive models are the ones which produce the best predictive model.

Significance tests were performed comparing the performance of user-adaptive models with the three algorithms and the dummy regressor. Results from the user-independent model with random forests also went through the significance test. The details can be found below:

Algorithm	Type pf system	t-statistics	pvalue
Random forests	user-adaptive	-2.192	0.0292
Random forests	user-independent	0.505	0.6162
GBM	user-adaptive	**-2.764**	**0.0061**
Ridge Regression	user-adaptive	-2.375	0.0182

Table 5: significance test results on three algorithms

As is shown above, the 3 user-adaptive models achieve statistically better performances than the baseline, especially the gradient boosting machines. For the user-independent model with random forests, although the mean squared error with random forests is smaller than the dummy regressor, the difference in their performances didn’t pass the significance test.

6.1.3 Selecting Sampling Methods

In this subsection, a comparison between different sampling methods was done. There are 3 options: use the whole conversation to compute features, in other words each sample in the feature set corresponds to the full conversation; use the first 10 seconds of the conversation as samples; use roughly the last 10 seconds of the conversation as samples. Here we have used a user-independent model to compare the sampling methods.

Algorithm	MSE	R_squared	Baseline_MSE	Baseline_R_squared	pvalue
Whole conversation as sample	**0.696**	**0.40**	**1.233**	**-0.09**	**0.0061**
First 10 seconds as sample	0.995	0.14	1.194	-0.03	0.3526
Last 10 seconds as sample	**0.713**	**0.36**	**1.259**	**-0.10**	**0.0056**

Table 6: Mean performances of GBM with different sapling methods

It can be seen that taking the whole conversation as samples performs better than taking segments from conversations as samples.

Readers might find it unreasonable that the baselines of the three sampling methods were different. This is due to the fact that when different sampling methods are taken, there will be unequal number of NAs appearing in the feature set, hence the samples remaining in the data set are also different.

6.2 Predicting Other Per-Conversation Tasks

After the series of experiments in the model building procedure with friendliness prediction, we settle down on the options which empirically generate better outputs. Therefore, the same options will also be chosen for other per-conversation tasks, in order to reduce the complexity of data analysis.

We employed the same pipeline to predict the emotional states of the participant, and the results are shown below:

	Top performance of the pipeline		Baseline		Significance test	
Task	MSE	R_squared	MSE	R_squared	t-statistic	p-value
Friendliness	**0.696**	**0.40**	**1.233**	**-0.09**	**-2.764**	**0.0061**
Arousal	3.187	0.45	3.628	0.54	-0.796	0.4266
Valence	1.994	0.27	2.741	0.35	-1.787	0.0750
Level of frustration	1.209	0.14	1.120	0.160	0.415	0.6785

Table 7: Comparison among different sampling methods

There is no model to be found except for the prediction of friendliness which outperforms the baseline significantly. The prediction of valence scores seems to be the only one promising. The intuitive explanation for such a phenomenon is that the perceived friendliness of the pedestrian influences the participant's impression of the conversation, and therefore influences the valence of the participant's emotion.

The failure in predicting other per-conversation scores is undesirable yet understandable, based on the previous analysis and comparison between participants' rating and annotators' ratings. Those per-conversation scores were self-reported by the participants after each conversation, meaning it could suffer from casual scoring, different understanding of social signals, as well as different standards for scoring. This explanation is also supported by the user study where 3 annotators were recruited to rate the friendliness of the conversation (5.4.2 (a)).

6.3 Predicting Per-Person Tasks

Unlike per-conversation predictions, the per-person prediction has a much smaller sample size, since there are only one target value for each participant. Therefore, we have chosen a leave-one-subject approach to train and evaluate the models.

6.3.1 Predicting personality

Additional preprocessing was applied on the feature set in order to have the per-person feature-target samples. The mean of the features of all conversations from the same participant was calculated to replace the original feature set. Consequently, the data set only has 20 records, where each record corresponds to the nonverbal conversational behaviors recorded with the same participant.

Owing to the small size of the data set for per-person situations, the data set itself as well as the model could contain much variance. The results can be found below:

Task	MSE_test	MSE_baseline	t-statistic	pvalue
Extraversion	106.50	100.31	0.233	0.820
Neuroticism	69.08	46.77	0.158	0.880
Conscientiousness	56.90	32.41	0.771	0.445
Agreeableness	70. 69	54.79	0.440	0.663
Openness to experience	59.68	50.48	0.479	0.635

Table 8: Top performances in personality prediction

As expected, no predictive model superior to the baseline can be trained with the data set.

6.3.2 Predicting racial prejudice

The prediction of racial prejudice is different from that of personality due to the reason that features which depicts the difference between the conversations with Arabians and Caucasians should be computed and used as the predictors. Therefore, we have generated another set of features based on the original feature set of the conversations with Arabians and Caucasians. Specifically, we have subtracted the mean of the features of "Caucasians" from that of the "Arabians". The performance of the predictive model of implicit and explicit racial prejudice are as follows:

Task	MSE_test	MSE_baseline	t-statistic	p-value
Implicit racial prejudice	0.276	0.224	0.957	0.3468
Explicit racial prejudice	6411.82	2264.26	0.194	0.8472

Table : performance of predicting implicit and explicit racial prejudice

Likewise, the model also fails to generate predictions about people' racial bias better than the dummy regressor.

6.4 Summary

Summarising the results from the above subsections, a few tentative conclusions have been drawn:

- For self-report per-conversation tasks, the scores are greatly influenced by the unfavorable factors during the study, such as individuals' differences in the standards for friendliness rating, the unwanted delay of the user to report the scores (unexpected delay in filling the experience report), or casual rating from the participants. Therefore, there's no good enough predictive models generated for these topics. To have more powerful predictive models, one option is to have a third person to rate the conversations in a standardized way. The improvement of such actions was already explained in section 6.1.

- For the per-person tasks, the disadvantage comes from the small number of sample size. Compared to the a feature size of more than 100, the sample size for per-person predictions is considerably small: 20. However, it can be expected that the situation will improve as the data set size increases.

Chapter 7 Reflection and Future Work

In this section, the imperfections in the whole study will be explained and provided with a possible solution in two perspectives.

7.1 User Study

Much efforts have been put into the user study stage of the whole project. Details about the questionnaires, the experiment instructions, the sequence of each step, as well as the cover story were considered and discussed about multiple times. The major obstacles we met during the user study stage are participant recruitment, hardware set-up and the difficulty in making the participants to follow the procedure.

The obstacle that directly relates to the quality of the data set was in hardware set-up. Since it was a multimodal framework for capturing and recording the social signals from people, the task was to assure the devices will be running for more than 2 hours in different environments such as strong direct sunlight or rainy weathers. Although much considerations were taken, the following technical issues still appeared occasionally either in the early stage of the study, or even throughout the whole study:

a. Abrupt system failure, which leads to a completely useless recording, especially in the early stage of the user study
b. Failure of a certain channel in some cases, which leads to missing values or abnormal values in the extracted features. For instance, overexposure in strong sunlight leads to the failure of the system to detect eye movements and gaze location.

If the above mentioned obstacles were solved, there could exist a larger data set with more robust and correct features, making a predictive model more feasible.

Furthermore, the mobile setting has invited much unpredictability into the study. The noise level of the environment, the weather, the number of passengers in the area, as well the personalities of the pedestrians are all uncontrollable factors in the study. To process the dynamics from the mobile settings successfully, the future work could be to think of noise removal methods. They could mean:

a. Utilizing the hardware in the correct and most suitable way so as to increase the recording precision and quality
b. The location of the study should avoid too crowded or noisy areas while reserving the in-the-field fashion
c. Implement tools to remove the noise afterwards, such as stabilize the recorded videos before doing feature extraction from the video

Finally, as has been investigated in the previous chapter, the predictive models showed distinctive contrast in performance. Since the scores such as friendliness are self-report ones, and people's ways of receiving and understanding signals as well as people's standards for giving scores, the prediction could be difficult. One thing to do is to write descriptions for each rating scales, therefore the participants will have a definite and clearer understanding about the criteria for ratings. Another option could be to process on the original scores. To deal with the two sources of rating variations: shift of average ratings and different rating scales, a decoupling normalization method could be used [79].

7.2 Data Process

7.2.1 Intermediate Checks in the Pipeline

The data process stage of the study involves much coding work. It needs to be assured that the every step of the processing should have zero mistake. It happened several times that an error had been detected in the very final stage of a pipeline, or the performance of the predictive models were not up to expectations, resulting in a thorough rework. One thing to solve the issue is to check at the intermediate stages, preferably by printing out the values one need to check about. For instance, for the feature "smile", a few tests can be done to evaluate it:

- For the videos recorded with the world camera of Pupil Labs Headset, after the facial action unit analysis with openSMILE, the level of this feature can be printed onto the video to demonstrate the real-time level of smile as detected by openSMILE. In this case the experimenter could examine manually whether the value correctly reflects the intensity of smile. This also applies for the second-person-view gaze estimation. With the gaze estimation model, for every frame of the video, whether or not the pedestrian is looking at the face of the participant will be given. One can ploy this value on the same video and check manually about the accuracy of the model.
- One thing to check whether the smile were computed correctly is to test whether the correlation between the feature and the target value is in line with hypothesis. For instance, it was expected that the intensity of smile should be positively correlated with friendliness level of the person. In this case, correlation test can be done to verify if such a pattern exist.

7.2.2 Influence of the Sample Size

In this subsection, further analysis will be done to discuss about and prove the possibility of improvement and future work. Below is a graph showing the relationship between mean squared error and the size of recordings in the training set in the task of friendliness prediction. The x axis corresponds to the number of participants whose recordings are utilized in the training set, and the y axis corresponds to the mean performance.

It can be seen that as the number of participants (and hence conversations) increases in the training set, the performance of the dummy regressor remains roughly unchanged, while the performance of the support vector regressor constantly rises.

Therefore it can be implied from the above graph that when the size of the training set increases, the current pipeline should be able to output better predictive models. This has demonstrated the importance of increasing the participant numbers in the study.

7.2.3 Manipulation with Feature Set

In the additional user study (appendix 6), annotators have pointed out the dual meanings of a same action. For example, when the pedestrian casts his or her eyes off the face of the participant, two opposite causes can be given: the pedestrian was impatient in the conversation therefore was easily distracted by the environment; or the pedestrian was indeed thinking over some suggestions from the participant and therefore looked at other directions during consideration. Similar conditions also appear for actions such as smile, which could be either supportive and positive, or negative and sarcastic.

Finally, the randomness that comes with the mobile setting should be dealt with. One example is that when the pedestrian is detected to be frowning, the reason could be other than emotional expression. When the environment is sunny during the conversations, there's high chance that the pedestrian is facing too strong sunlight, therefore leading to a facial expression of frowning. The system is unable to tell the cause of such an action, therefore interpret it as a signal of negative emotion.

The user study has also revealed the difficulties in such a predictive model. As is summarized in the study, 9 out of the 17 clues the annotators have marked as important in determining the friendliness of the subject were not or unable to be implemented. The major problem with those features is that unlike actions such as frowning or smiling, many activities such as taking a book out of a bag during a conversation can not be detected and coded with a relatively low level of artificial intelligence. However, such actions of escaping from the conversation were noted to be very important clues in determining friendliness. Those features remain for higher level of artificial intelligent systems to solve.

Reference

[1, 36] Vinciarelli, A., Pantic, M., & Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. Image and vision computing, 27(12), 1743-1759.

[2] Lippmann, W. (1946). *Public opinion*. Transaction Publishers.

[3] Nelson, T. D. (Ed.). (2009). Handbook of prejudice, stereotyping, and discrimination. Psychology Press.

[4] Sue, D. W. (2001). Multidimensional facets of cultural competence. The counseling psychologist, 29(6), 790-821.

[5, 17, 20, 22] Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. Journal of personality and social psychology, 82(1), 62.

[6] Dovidio, J. F., Kawakami, K., Johnson, C., Johnson, B., & Howard, A. (1997). On the nature of prejudice: Automatic and controlled processes. Journal of experimental social psychology, 33(5), 510-540.

[7] Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. European journal of social psychology, 25(1), 57-75.

[8] Vanman, E. J., Paul, B. Y., Ito, T. A., & Miller, N. (1997). The modern face of prejudice and structural features that moderate the effect of cooperation on affect. Journal of personality and social psychology, 73(5), 941.

[10] Brigham, J. C. (1993). College students' racial attitudes. Journal of Applied Social Psychology, 23(23), 1933-1967.

[11, 70] Modern racism, ambivalence, and the Modern Racism Scale. McConahay, John B. Dovidio, John F. (Ed); Gaertner, Samuel L. (Ed). (1986). Prejudice, discrimination, and racism , (pp. 91-125). San Diego, CA, US: Academic Press, xiii, 337 pp.

[12] Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. Journal of personality and social psychology, 74(6), 1464.

[13] Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. Journal of experimental psychology: general, 106(3), 226.

[14] Kawakami, K., Dion, K. L., & Dovidio, J. F. (1998). Racial prejudice and stereotype activation. Personality and Social Psychology Bulletin, 24(4), 407-416.

[15] Hing, L. S. S., Li, W., & Zanna, M. P. (2002). Inducing hypocrisy to reduce prejudicial responses among aversive racists. Journal of Experimental Social Psychology, 38(1), 71-78.

[16] Nosek, B. A., & Banaji, M. R. (2001). The go/no-go association task. Social cognition, 19(6), 625-666.

[18] Fairbairn, C. E., Sayette, M. A., Levine, J. M., Cohn, J. F., & Creswell, K. G. (2013). The effects of alcohol on the emotional displays of Whites in interracial groups. Emotion, 13(3), 468.

[19] Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. Psychological review, 107(1), 101.

[21, 24, 28] Crosby, F., Bromley, S., & Saxe, L. (1980). Recent unobtrusive studies of Black and White discrimination and prejudice: A literature review. Psychological Bulletin, 87(3), 546.

[23] Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. Journal of personality and social psychology, 84(4), 738.

[25] Kawakami, K., Williams, A., Sidhu, D., Choma, B. L., Rodriguez-Bailón, R., Cañadas, E., ... & Hugenberg, K. (2014). An eye for the I: Preferential attention to the eyes of ingroup members. Journal of personality and social psychology, 107(1), 1.

[26] Knapp, M. L., Hall, J. A., & Horgan, T. G. (2013). Nonverbal communication in human interaction. Cengage Learning.

[27] Harrigan, J. A., Wilson, K., & Rosenthal, R. (2004). Detecting state and trait anxiety from auditory and visual cues: A meta-analysis. Personality and Social Psychology Bulletin, 30(1), 56-66.

[28] Burgoon, J. K., Blair, J. P., & Strom, R. E. (2008). Cognitive biases and nonverbal cue availability in detecting deception. Human Communication Research, 34(4), 572-599.

[29, 32] Palazzi, A., Calderara, S., Bicocchi, N., Vezzali, L., di Bernardo, G. A., Zambonelli, F., & Cucchiara, R. (2016, September). Spotting prejudice with nonverbal behaviours. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 853-862). ACM.

[30] Hinde, R. A. (1972). Non-verbal communication. Cambridge University Press.

[31] Dovidio, J. F., Pearson, A. R., & Orr, P. (2008). Social psychology and neuroscience: Strange bedfellows or a healthy marriage?. Group Processes & Intergroup Relations, 11(2), 247-263.

[33, 68] Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. Journal of personality and social psychology, 59(6), 1216.

[34] Hathaway, S. R., & McKinley, J. C. (1951). Minnesota Multiphasic Personality Inventory; Manual, revised.

[35] Eysenck, S. B., Eysenck, H. J., & Barrett, P. (1985). A revised version of the psychoticism scale. Personality and individual differences, 6(1), 21-29.

[37] Scherer, K. R. (1979). Personality markers in speech. Cambridge University Press.

[38] Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. Journal of Personality and Social Psychology, 37(5), 715.

[39] Scherer K.R. Personality markers in speech. In Scherer K.R. and Giles H. (eds.) Social Markers in Speech, pp. 147-209. Cambridge University Press, 1979.

[40] Polzehl, T., Moller, S., & Metze, F. (2010, September). Automatically assessing personality from speech. In Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on (pp. 134-140). IEEE.

[41] Pianesi, F., Mana, N., Cappelletti, A., Lepri, B., & Zancanaro, M. (2008, October). Multimodal recognition of personality traits in social interactions. In Proceedings of the 10th international conference on Multimodal interfaces (pp. 53-60). ACM.

[42] Larsen, R. J., & Shackelford, T. K. (1996). Gaze avoidance: Personality and social judgments of people who avoid direct face-to-face contact. Personality and individual differences, 21(6), 907-917.

[43] Broz, F., Lehmann, H., Nehaniv, C. L., & Dautenhahn, K. (2012, September). Mutual gaze, personality, and familiarity: Dual eye-tracking during conversation. In RO-MAN, 2012 IEEE (pp. 858-864). IEEE.

[44] Hoppe, S., Loetscher, T., Morey, S., & Bulling, A. (2015, September). Recognition of curiosity using eye movement analysis. In Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (pp. 185-188). ACM.

[45] Walters, M. L., Dautenhahn, K., Te Boekhorst, R., Koay, K. L., Kaouri, C., Woods, S., ... & Werry, I. (2005, August). The influence of subjects' personality traits on personal spatial zones in a human-robot interaction experiment. In Robot and Human Interactive Communication, 2005. ROMAN 2005. IEEE International Workshop on (pp. 347-352). IEEE.

[46] Smith, C. A., & Ellsworth, P. C. (1985). Patterns of cognitive appraisal in emotion. Journal of personality and social psychology, 48(4), 813.

[47] Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. Psychological bulletin, 115(1), 102.

[48] Fasel, B., & Luettin, J. (2003). Automatic facial expression analysis: a survey. Pattern recognition, 36(1), 259-275.

[49] Schuller, B., Rigoll, G., & Lang, M. (2004, May). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on (Vol. 1, pp. I-577). IEEE.

[50] Nogueiras, A., Moreno, A., Bonafonte, A., & Mariño, J. B. (2001). Speech emotion recognition using hidden Markov models. In Seventh European Conference on Speech Communication and Technology.

[51, 73, 74] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.

[52] Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology, 45(4), 602-607.

[53] Wagner, J., Lingenfelser, F., Baur, T., Damian, I., Kistler, F., & André, E. (2013, October). The social signal interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In Proceedings of the 21st ACM international conference on Multimedia (pp. 831-834). ACM.

[54] Camurri, A., Coletta, P., Massari, A., Mazzarino, B., Peri, M., Ricchetti, M., ... & Volpe, G. (2004). Toward real-time multimodal processing: EyesWeb 4.0. In Proc. Artificial Intelligence and the Simulation of Behaviour (AISB) 2004 Convention: Motion, Emotion and Cognition (pp. 22-26).

[55] McDuff, D., Mahmoud, A., Mavadati, M., Amr, M., Turcot, J., & Kaliouby, R. E. (2016, May). AFFDEX SDK: A Cross-Platform Real-Time Multi-Face Expression Recognition Toolkit. In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 3723-3726). ACM.

[56] Littlewort, G., Whitehill, J., Wu, T., Fasel, I., Frank, M., Movellan, J., & Bartlett, M. (2011, March). The computer expression recognition toolbox (CERT). In Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on (pp. 298-305). IEEE.

[57] OpenFace: an open source facial behavior analysis toolkit Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, in IEEE Winter Conference on Applications of Computer Vision, 2016

[58] Dalmaijer, E.S., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments. Behavior Research Methods, 46, 913-921. doi:10.3758/s13428-013-0422-2

[59] Paul Boersma & David Weenink (2013): Praat: doing phonetics by computer [Computer program]. Version 5.3.51, retrieved 2 June 2013 from http://www.praat.org/

[60] Poria, S., Chaturvedi, I., Cambria, E., & Hussain, A. (2016). Convolutional MKL based multimodal emotion recognition and sentiment analysis. ICDM. Barcelona.

[61] Kim, Y., Lee, H., & Provost, E. M. (2013, May). Deep learning for robust feature generation in audiovisual emotion recognition. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 3687-3691). IEEE.

[62] Bousmalis, K., Mehu, M., & Pantic, M. (2009, September). Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (pp. 1-9). IEEE.

[63] machine intelligence 31.1 (2009): 39-58. Zeng, Z., Pantic, M., Roisman, G. I., & Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1), 39-58.

[64, 66] Palazzi, A., Calderara, S., Bicocchi, N., Vezzali, L., di Bernardo, G. A., Zambonelli, F., & Cucchiara, R. (2016, September). Spotting prejudice with nonverbal behaviours. In Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (pp. 853-862). ACM.

[65] Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and using the Implicit Association Test: II. Method variables and construct validity. Personality and Social Psychology Bulletin, 31(2), 166-180.

[67] Echebarria-Echabe, A., & Guede, E. F. (2007). A new measure of anti-Arab prejudice: Reliability and validity evidence. Journal of Applied Social Psychology, 37(5), 1077-1091.

[69] Park, J., Felix, K., & Lee, G. (January 01, 2007). Implicit Attitudes Toward Arab-Muslims and the Moderating Effects of Social Information. Basic and Applied Social Psychology, 29, 1, 35-45.

[71] Yik, M. S., Russell, J. A., & Barrett, L. F. (1999). Structure of self-reported current affect: Integration and beyond. Journal of personality and social psychology, 77(3), 600.

[72] Florian Eyben, Felix Weninger, Florian Gross, Björn Schuller: “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor”, In Proc. ACM Multimedia (MM), Barcelona, Spain, ACM, ISBN 978-1-4503-2404-5, pp. 835-838, October 2013. doi:10.1145/2502081.2502224

[75] Bulling, A., Ward, J. A., Gellersen, H., & Troster, G. (2011). Eye movement analysis for activity recognition using electrooculography. IEEE transactions on pattern analysis and machine intelligence, 33(4), 741-753.

[76] Ekman, P., & Friesen, W. V. (1977). Facial action coding system.

[77] Zhang, X., Sugano, Y., & Bulling, A. (2017). Everyday Eye Contact Detection Using Unsupervised Gaze Target Discovery.

[78] Zhang, X., Sugano, Y., Fritz, M., & Bulling, A. (2016). It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. arXiv preprint arXiv:1611.08860.

[79] Jin, R., & Si, L. (2004, July). A study of methods for normalizing user ratings in collaborative filtering. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 568-569). ACM.

Appendix 1: Introduction and Instruction

Introduction

In this study, your task is to recruit study participants for us. They have to meet certain criteria listed below. While you are recruiting participants for our studies, you will wear video and audio recording equipment. You will also complete a short questionnaire about your affective state after each interaction. With this, we want to investigate, how people's affective state is influenced by social interactions.

You are supposed to recruit participants from the whole campus area, but only from the following two groups: 1) Europeans/Americans and 2) Arabs. We ask you to recruit participants from these two groups in equal parts. The reason is, that we want to make use of those groups to investigate intercultural aspects of affect response and discussion behaviour, when they take part in our experiments.

The information of the two experiments are provided below:

a. *The first task takes around 2.5 hours. The participant will wear a microphone, a camera and an eye tracker, walk around in the campus and have conversations with people. The task will be to recruit people from the street as new participants in the experiment. (same as the current experiment)*

b. *The second task takes 1 hour. The participant will have conversations with several other participants in a room. A part of the conversation will be recorded, in order to build an automatic analysis system for discussions.*

Instructions

Task:

Recruit as many **strangers** as possible for two experiments from 2 groups: Europeans/Americans, and Arabs. Try to keep the number of people recruited balanced, across **ethnic groups** and **genders**.

Process:

- Before the formal conversation:

  Check if the pedestrian meets the requirements: 1) **a stranger for you**, 2) **speaks fluent German**, 3) **Europeans/Americans or Arabs**.

- During the conversation:
  - Ask for the permission to talk to the pedestrian for a few minutes.

- Introduce about the two projects in your own words: e.g. "*Currently I am recruiting participants for two experiments about conversational behaviours.*

  <Your descriptions of two experiments>

  Registration for both of the experiments is more than welcomed. We will contact you separately for each experiment as long as the requirements are met."
- If asked, explain more about the experiments, or even about the on-body hardware.

- If your subject confirms that he/she has time and are committed to participating in the experiment, go to step 1. Otherwise if your subject refuse to take part in our further experiments, go directly to step 2. The process is as follows:

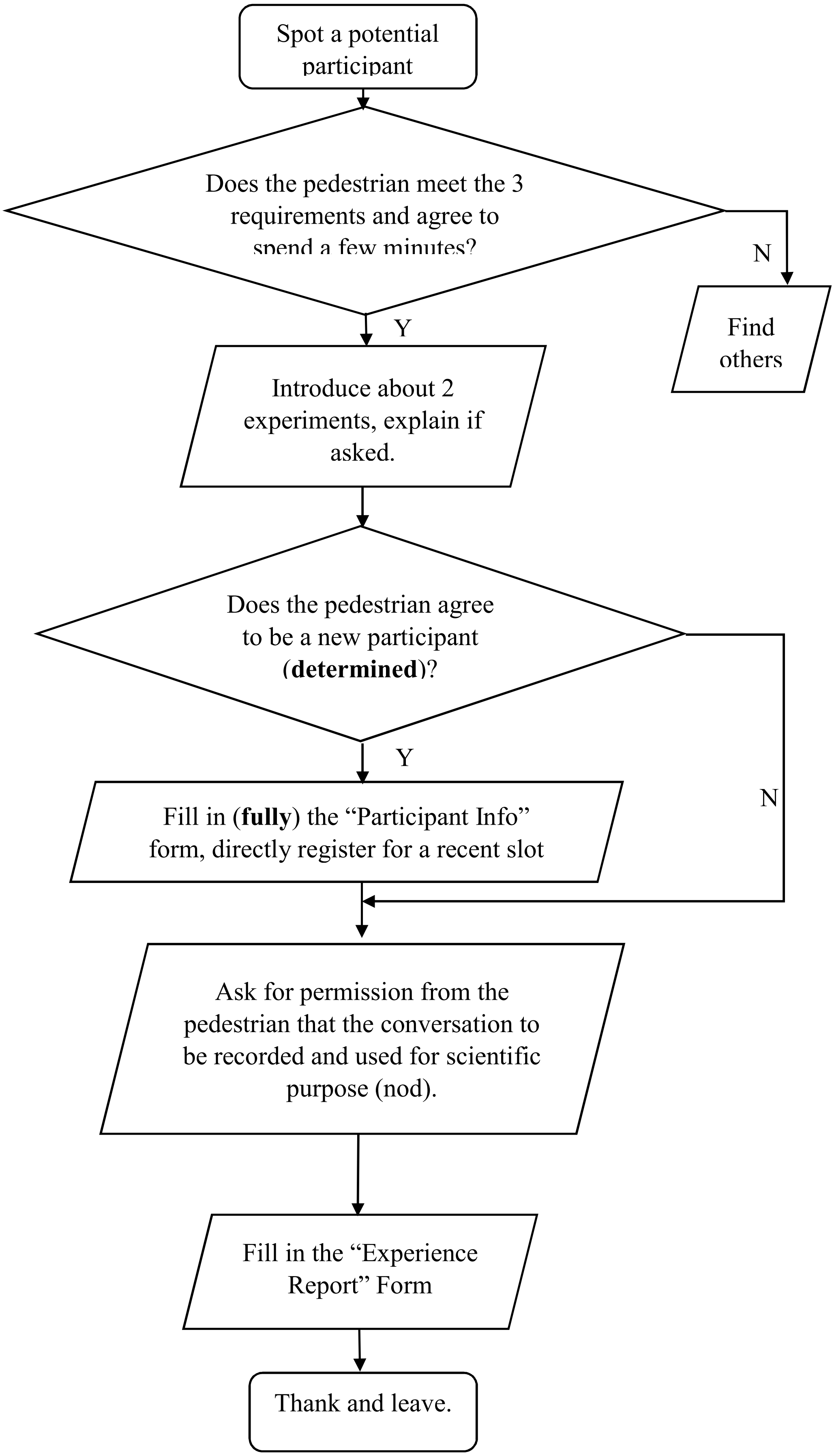
  1. Ask for the subject's personal information, including name, age, gender, student number, study of subject, phone number, email address and nearest available time in the next 7 days. **Fill the information in a form** (Participant Info) and directly register for a recent available slot via doodle link if possible. Then go to step 2.
  2. Ask the subject to nod if he/she agrees that the previous conversation be recorded and used for scientific purpose. e.g. "Do you agree that our conversation to be recorded with microphone and cameras, and the data to be used for scientific purposes and published in anonymised form in scientific publications? If yes, please nod to me."

➢ After the Conversation:

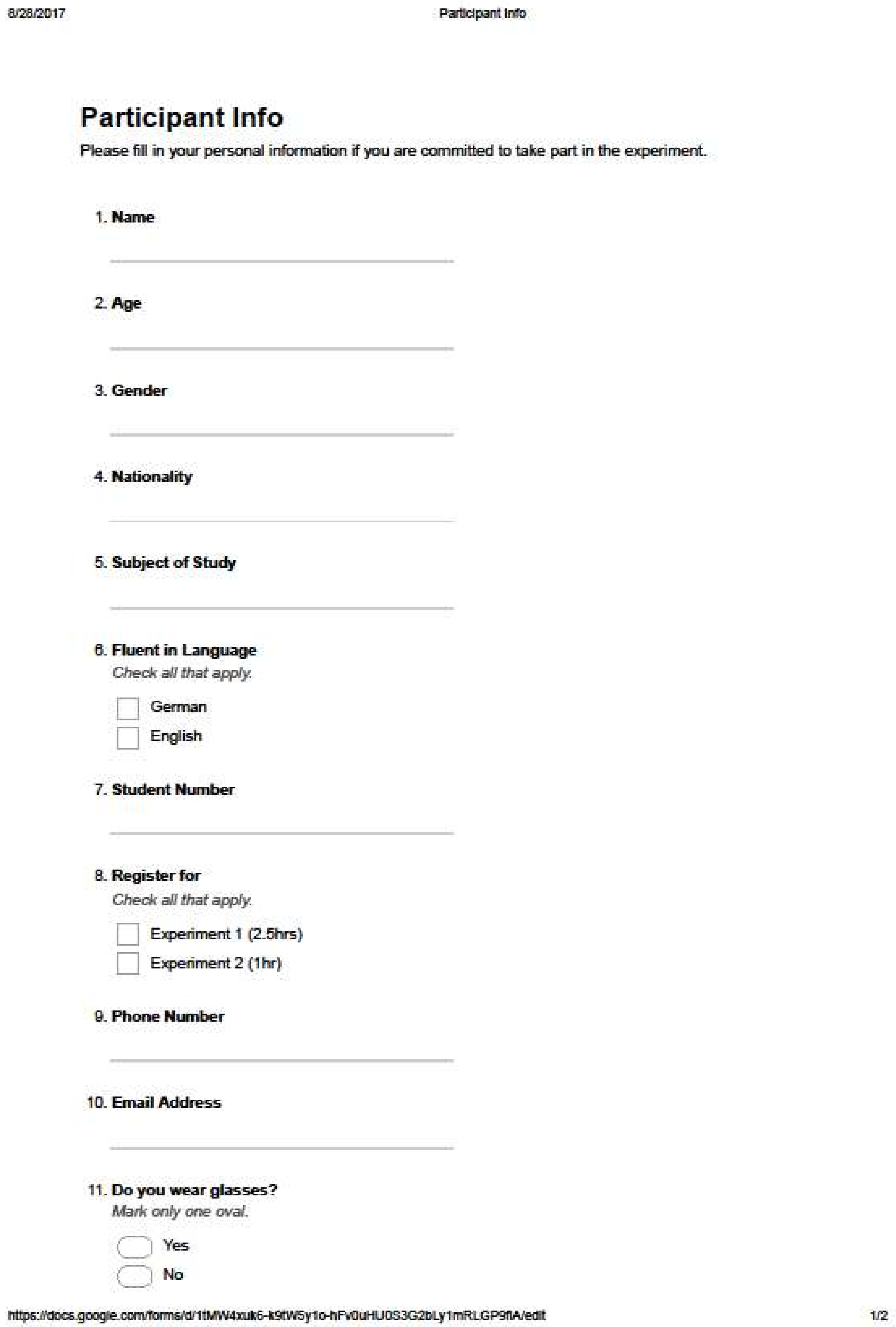
Fill in the **questionnaire** (Experience Report) about your experience in the previous conversation, including whether the person agreed to be a participant, your perceived friendliness of the person, and how upset you are after the conversation, etc.,

Tips for recording:

- **Don't move** the eye tracker during the recording.
- The recording requires good lighting and sound conditions. Please try to **avoid** too dark or too noisy places or direct sunlight.
- The recording requires **one-to-one** conversations, avoid talking to a group.
- To reduce the complexity of our recording, please only talk to people who are standing or walking.


Spot a potential participant
Does the pedestrian meet the 3 requirements and agree to spend a few minutes?
N
Find others
Y
Introduce about 2 experiments, explain if asked.
Does the pedestrian agree to be a new participant (**determined**)?
Y
N
Fill in (**fully**) the "Participant Info" form, directly register for a recent slot
Ask for permission from the pedestrian that the conversation to be recorded and used for scientific purpose (nod).
Fill in the "Experience Report" Form
Thank and leave.

Appendix 2: Personal Information Form



Participant Info

Please fill in your personal information if you are committed to take part in the experiment.

1. **Name**

2. **Age**

3. **Gender**

4. **Nationality**

5. **Subject of Study**

6. **Fluent in Language**
   Check all that apply.
   □ German
   □ English

7. **Student Number**

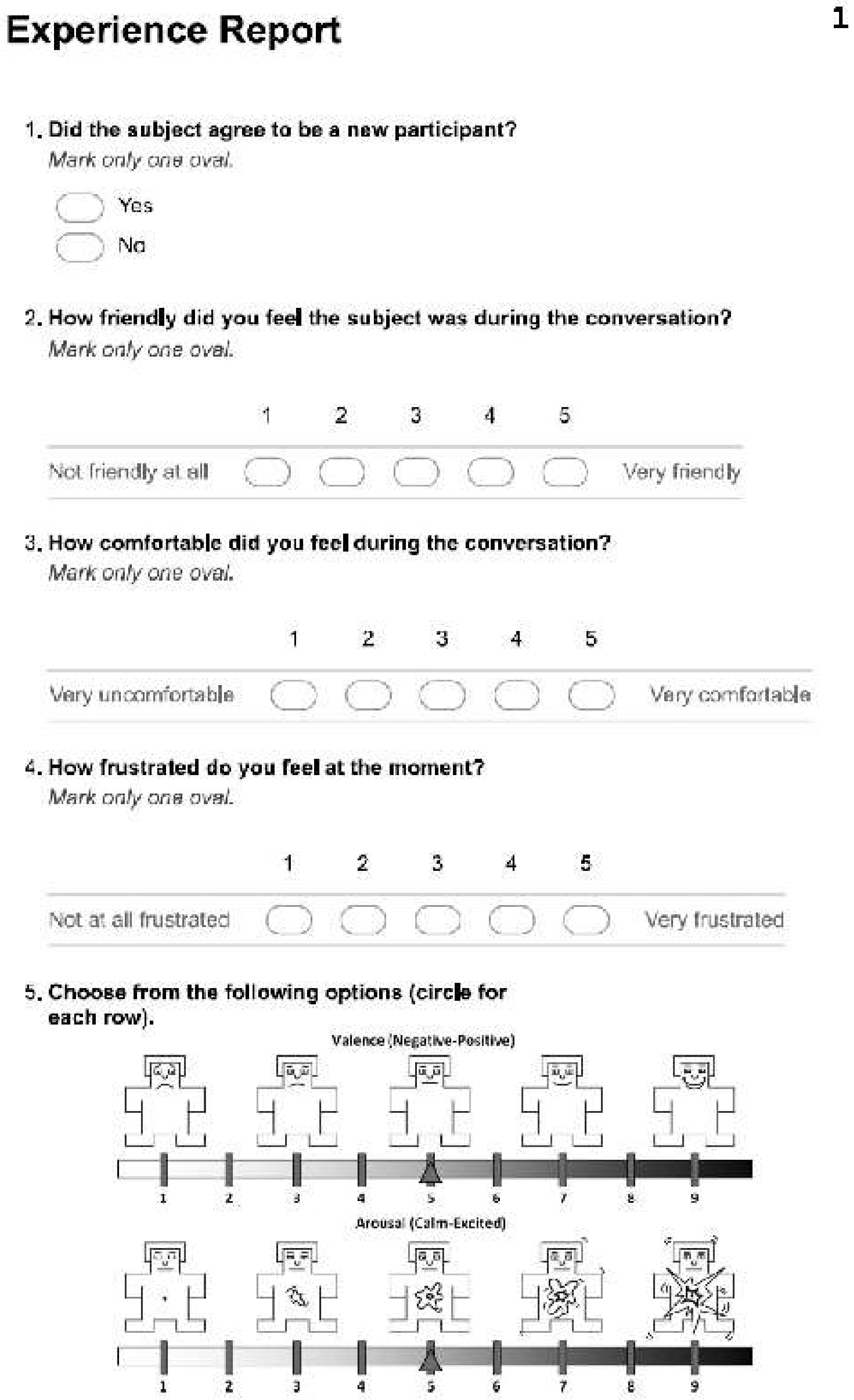
8. **Register for**
   Check all that apply.
   □ Experiment 1 (2.5hrs)
   □ Experiment 2 (1hr)

9. **Phone Number**

10. **Email Address**

11. **Do you wear glasses?**
    Mark only one oval.
    ○ Yes
    ○ No

Appendix 3: Experience Report



Experience Report

1

1. **Did the subject agree to be a new participant?**
Mark only one oval.

○ Yes
○ No

2. **How friendly did you feel the subject was during the conversation?**
Mark only one oval.

	1	2	3	4	5	
Not friendly at all	○	○	○	○	○	Very friendly

3. **How comfortable did you feel during the conversation?**
Mark only one oval.

	1	2	3	4	5	
Very uncomfortable	○	○	○	○	○	Very comfortable

4. **How frustrated do you feel at the moment?**
Mark only one oval.

	1	2	3	4	5	
Not at all frustrated	○	○	○	○	○	Very frustrated

5. **Choose from the following options (circle for each row).**

Valence (Negative-Positive)

1 2 3 4 5 6 7 8 9

Arousal (Calm-Excited)

1 2 3 4 5 6 7 8 9

Appendix 4: New Anti-Arab Attitudes Scale

	Strongly Disagree	Moderately Disagree	Slightly Disagree	Neither agree nor disagree	Slightly Agree	Moderatel y Agree	Strongl y Agree
1. Unsere Vorfahren haben nicht gegen Türken und Araber gekämpft, nur damit wir ihnen Europa überlassen.							
2. Der Islam ist eine archaische Religion, die sich nicht an unsere heutige Zeit anpassen kann.							
3. Der Islam respektiert die Menschenrechte.							
4. Eine Trennung zwischen Religion und Staat ist in der muslimischen Kultur unmöglich.							
5. Der Islam ist eine Gefahr für Frauen.							
6. Europa sollte den Islam als eine wichtige Religion anerkennen.							
7. Die Europäischen Staaten sollten die Kontrolle arabischer Immigranten verstärken.							
8. Im Herzen der meisten Araber befinden sich Hass gegen den Westen und der Jihad.							
9. Die meisten arabischen Länder							

sind fanatisch, nationalistisch, und stehen im Konflikt mit Menschenrechten.							
10. Wir müssen dem Risiko der Islamisierung mit einer Stärkung unserer christlichen Identität begegnen.							
11. Der Islam ist radikal und intolerant.							
12. Im Angesicht der Immigration von Muslimen (Türken. Algerier, Marokkaner, etc.) und ihrer hohen Geburtenrate, besteht in Europa das Risiko der Islamisierung.							
13. Arabische Immigranten sind sehr häufig in Verbrechen verwickelt.							
14. Arabische Immigranten sind Ballast für unsere Sozialsysteme.							
15. Araber sind alle gleich. Sie sind dem Westen gegenüber feindselig.							
16. Arabische Immigranten sind							

eine Bedrohung für das Gesundheitswesen (AIDS, tuberculosis, hepatitis, etc.).							
17. In städtischen Gebiten mit hohem Anteil von arabischen Immigranten sind Verbrechen häufiger.							
18. Araber haben zur europäischen Kultur und Wissenschaft beigetragen.							
19. Die Jahrhunderte der muslimischen Okkupation Spaniens waren eine kulturelle und ökonomische Blüteperiode.							
20. Die westliche Kultur ist der muslimischen Überlegen.							
21. Der Islam respektiert Frauen.							
22. Versuche, arabische Immigranten in die europäische Kultur zu integrieren, sind Zeitverschwendung.							
23. Um akzeptiert zu werden, müssen arabische Immigranten das Versprechen abgeben, sich unserer Kultur und unseren Gepflogenheiten anzupassen.							
24. Es ist inakzeptabel, dass Frauen in Europa den							

islamischen Schleier tragen.							
25. Die meisten Araber sind froh über Terrorismus, der den westlichen Interessen entgegenläuft.							
26. Araber sollten aufgrund ihres religiösen Fundamentalismus strikten Kontrollen unterliegen.							
27. Der Islam ist genaugenommen keine Religion, sondern eine terroristische Bewegung.							
28. Die Immigranten der zweiten Generation machen weiter, ohne sich in unsere Kultur zu integrieren, und behalten die Traditionen ihrer Eltern bei.							
29. Araber sind eine zukünftige Bedrohung für Europa.							
30. Araber sind unseren kulturellen Bezugspunkten gegenüber fremd (Rom und Griechenland).							
31. Araber lieben Frieden und Koexistenz.							
32. Der Islam ist eine große Religion und verdient unseren Respekt.							

33. Islam und Christentum teilen die selben universellen ethischen Prinzipien.							
34. Arabische Länder kümmern sich stärker um Bekehrung und dem Bauen von Moscheen in europäischen Ländern, als um arme Arabern.							
35. Der Islam predigt Toleranz, Respekt fuer den Menschen, und eine friedvolle Koexistenz aller Länder.							
36. Araber benutzen die europäische Demokratie, um ihre Kultur und ihre Sitten einzuführen.							
37. Die europäische Polizei sollte besonders viel Aufmerksamkeit auf arabische Immigranten legen, da diese eine echte Bedrohung für unsere Länder sind.							
38. Araber sind verdächtig, Terrorismus zu unterstützen. Sie müssen beweisen, dass sie pazifistisch sind.							
39. Araber, die unsere Kultur und Tradition							

nicht akzeptieren, müssen in ihre Länder zurück kehren.							
40. Viele Kapitel der Menschheitsgeschichte, die von Zivilisation und Toleranz handeln, wurden von Arabern geschrieben.							
41. Araber dürfen in Europa keinen Respekt erwarten, solange sie nicht die Christen in ihren Ländern respektieren.							
42. Araber sind hilfreicher und humanitärer als Leute aus dem Westen.							

Appendix 5: Consent Form

max planck institut
informatik

PERCEPTUAL
USER INTERFACES
perceptual.mpi-inf.mpg.de

Consent Form

Participant Name ____________________________

Title of Study: **Truth Mining from Nonverbal Behaviours in a Mobile Setting**

Researcher: Xin Jia
Perceptual User Interfaces Group
Max Planck Institute for Informatics
Campus E1 4, 66123 Saarbrücken
xinjia@mpi-inf.mpg.de +49(0) 1766 5679 401

We kindly ask you to read and understand the following explanation of the study's purpose and procedure.

Purpose of the Study

Our research intends to investigate patterns in people's nonverbal behaviours and figure out their relationships with the traits of people or outcomes of conversations. The potential topics include:

1) predicting personality of the participant from dyadic conversations
2) predicting racial bias of the participant from dyadic conversations
3) predicting response of pedestrians for experiment recruitment from dyadic conversations

A set of devices was used during the experiment: a realSense 3D camera, a Pupil eye tracker, and a microphone. The nonverbal behaviour recorded include the conversations, the first-person-view videos of the participant from both the eye tracker and the realSense 3D camera, as well as the eye movement data from the eye tracker.

To try and obtain unbiased or natural reactions and behaviours, we had to give you some false information at the beginning of the study. This was necessary for us to better understand how natural nonverbal behaviour could reflect the attitude or personality of participants, because early disclosure of the true research topic would have altered participants' behaviour, therefore making the predictive model invalid. We apologize for misleading you, but we believe this was the only way to examine the processes that are the objective of our research. In designing this study, we took care to minimize any possible risks or discomforts that might be related to the deception.

Confidentiality

Your identity as a participant will be kept strictly confidential. Only researchers involved in the project will be allowed access to the data. To ensure your anonymity, this consent form will be kept separate from the questionnaires and recordings at all times.

The data will only be used for research purposes and might be published – in full or in parts, in original or modified form – but always anonymised as part of scientific publications. Such publication may include the sensor data and the answers to the questionnaires.

To explain better about the considerations we have taken to protect your rights and privacy, here are a few steps we have done or planned:

1. The data types and contents we recorded are designed to reveal as little as possible about your identity. They include: your answers to a series of questionnaires and the implicit association tests; the video of your eye movements; your speech during the conversation; your egocentric view during the conversations (which in fact mainly depicts the behaviours of your confederates).

2. We will separate this consent form and your personal information with any data recorded about you. In other words, the recorded data will be anonymised.

Now that you understand the true nature of our study, you have the right to refuse the use of the data we collected from you for research purposes. You are free to ask us not to use your data in our study analysis. If you decline to let us use your data, you will still receive the payment just as we promised. This is entirely voluntary, but we hope to analyse as much data as possible to proceed with our study.

Because this experiment is ongoing, we request that you not share the true nature and purpose of this experiment with others who might potentially participate in our study.

If you agree to allow us to use the data, please sign this form below. You may keep a copy of this form for your future reference, if needed.

Thank you again for your participation in our research!

Payment
Participants will be paid 10 EUR/hour for participating in the experiment.

Risks
There are no risks to participants during the recording.

Further Information
To contact an independent person about this research please refer to:

Dr. Andreas Bulling
Head of the Perceptual User Interfaces Group
Max Planck Institute for Informatics
Campus E1 4, Saarbrücken
bulling@mpi-inf.mpg.de +49(0) / 681 9325 2128

Consent
I have had the opportunity to discuss this study and my questions have been answered to my satisfaction. I consent to allow the use of my data for research purposes.

Participant Signature ___________________________ **Date** __________

Investigator Signature ___________________________ **Date** __________

Appendix 6: Annotation Notes

This appendix enlists the clues that were noted as important for labeling friendliness level during manual annotation. Furthermore, the correlation of the scores given by annotators and participants are calculated and reported.

Three annotators were recruited to rate the friendliness of the pedestrians in 13 videos, without knowing the scores given by the participants. The videos cover 5 participants and all score levels, and were selected randomly. The annotators were only presented with the video from the egocentric camera.

Ratings range from 1 (very unfriendly) to 5 (very friendly). Moreover, the annotators are required to give their clues/basis of giving scores, in other words, find the important activities or factors for scoring friendliness. For instance, the subjects didn't smile at all in the video and was sometimes frowning. Based on that, the annotator gives a score of 2. Then he/she should take notes of the clues that have made used of.

In the following list, the clues that were implemented in the feature set will be noted with the feature name, while the rest will be given an explanation for no incorporation.

Summary of clues

1. Frequency and average intensity of smile (**AU6, AU12, smile, smile_peaks, peak_mean**)
2. Duration of the pedestrian looking at the camera (**p2,avg_p2_duration,percent2**)
3. Frequency of nodding
    - Nodding was not implemented since a motion detected in the scene could be because of motion of the participant or the motion of the pedestrian. Even if these two motions can be separated, to tell the difference between the pedestrian nodding or the pedestrian changing posture (thus leading to the face appearing up and down in the scene) isn't easy to implement.
4. Whether the pedestrian uses words to show agreement, e.g. "yes" "sure"
    - The feature was not included since the headworn microphone theoretically only captures the speech from the participant. In reality very low voice in poor quality of the pedestrian can be captured, but it doesn't meet the requirement for automatic detection of "yes" and "sure".
5. How often the pedestrian gives comments or questions
    - Similar reasons as 4
6. Whether the pedestrians shows posture of thinking/wondering, such as by touching face or quickly looking at certain directions
    - Gestures or head orientation of the pedestrian isn't easy to extract from a moving camera
7. How close the pedestrian is from the camera (**distance_mean**)
8. Duration of conversation (**length**)
9. Whether the pedestrian stops what he/she was working on and then focus on the conversation, e.g. stops eating or walking
    - This is a complex activity, thus can not be read automatically

10. Whether the subject start to work on other things during the conversation, e.g. tries to find irrelevant things in bag
    - Same as 9
11. How frequent the subject frowns (**frown**)
12. How frequent the person moves his/her body or change gestures, which shows impatience (**distance_variance**)
    - Others are the same as 6
13. Whether the pedestrian seems relaxed
    - Relaxed is a complicated condition which is a kind of impression combining different channels, can not be read readily
14. Whether the pedestrian gives fake smiles
    - To tell apart real and fake smiles involves higher level of computer vision and much work with dataset collection and model training
15. Whether the pedestrian raise eyebrow to show surprise (**AU1**)
16. Whether the pedestrian was easily distracted by passengers (**p2**)
17. How reluctant the person seems to agree to be interviewed. Does the person agree by saying "sure", or shrugged and agreed, or very quickly nods head
    - Same as 9

Taking a look at the scores given by the annotators and the participants, the correlations are as follows:

- the correlation between the scores given by the 3 annotators are 0.49, 0.35 and 0.55 (mean=0.46);
- The score correlation between the annotators and the participants are 0.58, -0.09, and 0.23 (mean=0.24).

The above study affirms the situation that different people have different ways of perception in friendliness and have different standards for giving ratings. Moreover, the difference between the experiences of on-the-spot conversation and video reviewing can not be overlooked. The result brings forward the possibility of disadvantages.