Chalearn LAP challenges on self-reported personality recognition and non-verbal behavior forecasting during social dyadic interactions: Dataset, design, and results

Palmero, Cristina; Barquero, German; Jacques Junior, Julio C. S.; Clapés, Albert; Núñez, Johnny; Curto, David; Smeureanu, Sorina; Selva, Javier; Zhang, Zejian; Saeteros, David; Gallardo-Pujol, David; Guilera, Georgina; Leiva, David; Han, Feng; Feng, Xiaoxue; He, Jennifer; Tu, Wei-Wei; Moeslund, Thomas B.; Guyon, Isabelle; Escalera, Sergio

Published in:
Understanding Social Behavior in Dyadic and Small Group Interactions

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

ChaLearn LAP Challenges on Self-Reported Personality Recognition and Non-Verbal Behavior Forecasting During Social Dyadic Interactions: Dataset, Design, and Results

Cristina Palmero CRPALMEC7@ALUMNES.UB.EDU
German Barquero GBARQUGA9@ALUMNES.UB.EDU
Universitat de Barcelona and Computer Vision Center, Spain

Julio C. S. Jacques Junior JJACQUES@CVC.UAB.CAT
Computer Vision Center, Spain

Albert Clapés ALCL@CREATE.AAU.DK
Aalborg University, Denmark, and Computer Vision Center, Spain

Johnny Núñez JNUNEZCA11@ALUMNES.UB.EDU
Universitat de Barcelona and Computer Vision Center, Spain

David Curto DAVID.CURTO@ESTUDIANTAT.UPC.EDU
Universitat Politècnica de Catalunya, Spain

Sorina Smeureanu SSMEURSM28@ALUMNES.UB.EDU
Javier Selva JSELVACA21@ALUMNES.UB.EDU
Zejian Zhang ZZHANGZH45@ALUMNES.UB.EDU
Universitat de Barcelona and Computer Vision Center, Spain

David Saeteros DAVID.SAETEROSP@UB.EDU
David Gallardo-Pujol DAVID.GALLARDO@UB.EDU
Georgina Guilera GGUILERA@UB.EDU
David Leiva DLEIVAUR@UB.EDU
Universitat de Barcelona, Spain

Feng Han HANFENG@4PARADIGM.COM
Xiaoxue Feng FENGXIAOXUE@4PARADIGM.COM
Jennifer He HEYUXUAN@4PARADIGM.COM
Wei-Wei Tu TUWEIWEI@4PARADIGM.COM
4Paradigm, China

Thomas B. Moeslund TBM@CREATE.AAU.DK
Aalborg University, Denmark

Isabelle Guyon GUYON@CHALEARN.ORG
LISN (CNRS/INRIA) Université Paris-Saclay, France, and ChaLearn, USA

Sergio Escalera SERGIO@MAIA.UB.ES
Universitat de Barcelona and Computer Vision Center, Spain

Abstract

This paper summarizes the 2021 ChaLearn Looking at People Challenge on Understanding Social Behavior in Dyadic and Small Group Interactions (DYAD), which featured two tracks, self-reported personality recognition and behavior forecasting, both on the UDIVA v0.5 dataset. We review important aspects of this multimodal and multiview dataset consisting of 145 interaction sessions where 134 participants converse, collaborate, and

compete in a series of dyadic tasks. We also detail the transcripts and body landmark annotations for UDIVA v0.5 that are newly introduced for this occasion. We briefly comment on organizational aspects of the challenge before describing each track and presenting the proposed baselines. The results obtained by the participants are extensively analyzed to bring interesting insights about the tracks tasks and the nature of the dataset. We wrap up with a discussion on challenge outcomes, and pose several questions that we expect will motivate further scientific research to better understand social cues in human-human and human-machine interaction scenarios and help build future AI applications for good.

Keywords: Personality recognition, Behavior forecasting, Multimodal approaches, Human interaction, AI competitions

1. Introduction

Socially intelligent systems are expected to accurately perceive, understand, react, and adapt to the affective and cognitive state of interacting individuals in different contexts so as to provide a more natural, empathic, tailored communication (Dautenhahn, 2007; Esposito et al., 2014; Paiva et al., 2017; Nocentini et al., 2019). To embody this human-likeness into such systems, it is crucial to have a deeper understanding of real human-human interactions first, to allow for the computational modeling of individual and social behaviors, and interpersonal influence (Burgoon et al., 1995; Escudero et al., 2018; Braithwaite and Schrodt, 2021). Research in dyadic and small group interactions has enabled the development of automatic approaches for detection, understanding, modeling, and synthesis of individual and interpersonal behaviors, social signals, and dynamics (Picard, 2000; Gatica-Perez, 2009; Vinciarelli et al., 2011, 2015). For measuring interpersonal processes during an interaction such as non-verbal synchrony (Delaherche et al., 2012), rapport (Zhao et al., 2016), or engagement (Dermouche and Pelachaud, 2019), the joint modeling of all interlocutors and/or other sources of context has been frequently considered. These sources of context may include individual factors, such as sociodemographics and other attributes of each interlocutor, or shared factors, such as the history of the interaction and characteristics of the situation (Rauthmann et al., 2014). However, for the task of recognizing individual attributes or behaviors in interaction settings, most computational approaches usually only consider information from the target interlocutor, disregarding the effect of any sources of context on individual behavior in addition to existing dyadic or group interdependencies (Barrett et al., 2011; Wright, 2014; Vinciarelli and Mohammadi, 2014b; Vinciarelli et al., 2015; Moore, 2017; Dudzik et al., 2019; Corona et al., 2020b).

To foster research on interlocutor- and context-aware approaches for social behavior modeling and understanding, we organized the *ChaLearn LAP Challenge on Understanding Social Behavior in Dyadic and Small Group Interactions* (henceforth referred to as DYAD'21 Challenge), in conjunction with the *International Conference on Computer Vision* (ICCV), 2021. In particular, the challenge was divided into two competition tracks, related to key elements for understanding social interactions and developing socially intelligent systems: 1) Self-reported Personality Recognition, with the goal of inferring the self-reported personality of a single individual based on the individual's behavior under different interaction-based situations; and 2) Behavior Forecasting, with the goal of predicting 2D hand, face and body landmarks of both interlocutors in an interaction up to two seconds in the future given an observed time window of at least four seconds. Both tracks were

based on the UDIVA v0.5 dataset, a subset of the UDIVA dataset (Palmero et al., 2021). The UDIVA v0.5 dataset consists of audiovisual, transcripts, and metadata information, as well as automatically extracted annotations, of 145 face-to-face dyadic interaction sessions. Such sessions are performed by 134 individuals of varied characteristics, and are divided into different conversational, collaborative, and competitive tasks.

We provided reference results from our own baseline architectures for each of the competition tracks. For self-reported personality, we used a Transformer-based architecture (Palmero et al., 2021) with audiovisual information from the target participant and the other interlocutor as context, as well as the associated individual and dyadic metadata. The challenge participants were able to outperform the provided baseline, but found the careful handling of metadata to be key for the task. In particular, they indicated gender and age to greatly influence the prediction. Challenge participants also found the conversation transcripts to be particularly useful for the task. For the behavior forecasting track, participants were not able to compete with a challenging but straightforward zero-velocity baseline. We attribute this to several factors: the stochasticity of hand gestures, the mostly static nature of seated social interactions (in comparison to highly interactive scenarios like dancing), the difficulty of accurately annotating hand poses, and the metric that evaluated predictions at frame level, which penalizes out-of-sync behaviors.

In this paper, we first describe the dataset used for the challenge (Section 2), as well as the challenge design and participation instructions and statistics (Section 3). Then, we describe both challenge tracks (Track 1: Self-Reported Personality Recognition in Section 4 and Track 2: Behavior Forecasting in Section 5), including a summary of the state of the art of each topic, baseline methodologies and solutions proposed by the top-performing teams, in addition to a discussion of the results obtained and the outcomes of each track. Furthermore, we discuss ethical considerations of the challenge data and potential implications stemming from research in personality recognition and behavior forecasting as the two main research topics considered herein (Section 6). Finally, Section 7 concludes the paper.

2. UDIVA v0.5 Dataset

In this section, we describe the UDIVA v0.5 dataset, which was used for both tracks of the DYAD'21 Challenge, and is currently publicly available for research purposes[1]. UDIVA v0.5 is a preliminary subset of UDIVA (Palmero et al., 2021), a highly varied multimodal, multiview dataset of zero- and previous acquaintance, face-to-face dyadic interactions, consisting of 180+ interaction sessions where 140+ participants arranged in dyads performed a set of different conversational, collaborative, and competitive tasks in a lab setting. UDIVA was collected using multiple audiovisual and physiological sensors, and includes sociodemographic, personality, internal state, and relationship metadata of all participants, as well as contextual information of the tasks. UDIVA v0.5 contains a subset of the interaction sessions, participants, synchronized camera views and metadata from UDIVA, in addition to new annotations collected for the challenge, which we detail below. For more information about the complete UDIVA dataset, we refer the reader to the work of Palmero et al. (2021).

1. https://chalearnlap.cvc.uab.cat/dataset/41/description/

2.1. Setup and structure of a dyadic session

Participants were recruited through university and social media ads and word of mouth. Prior to their first dyadic session, participants filled in a consent form in compliance with the EU General Data Protection Regulation (GDPR[2], European Parliament and Council of European Union 2016) as well as several questionnaires about themselves. UDIVA v0.5 includes information from self-reported sociodemographics (see participant's metadata in Section 2.4.2) and Big Five Inventory (BFI-2) personality questionnaires (Gallardo-Pujol et al., 2021). The latter provides personality ratings across five traits, known as OCEAN (see Section 4.1 for a description), in a 1-to-5 scale, which were later converted to z-scores using descriptive data from normative samples (Gallardo-Pujol et al., 2021). Before and after each interaction session, participants also completed several questionnaires regarding their current internal state. In particular, UDIVA v0.5 includes the pre- and post-session information from mood (Gallardo-Pujol et al. 2013, with items drawn from PEQPN by Williams et al. 2002, as a 1-to-5 rating scale) and fatigue (ad-hoc 1-to-10 rating scale) questionnaires. In particular, the mood questionnaire assessed current state based on eight classes: *Good*, *Bad*, *Happy*, *Sad*, *Friendly*, *Tense*, and *Relaxed*. Participants that did not fill in the pre- and/or post-fatigue questionnaire had their fatigue level set to 0.

To perform the administered tasks, participants were asked to sit at 90° relative to one another as depicted in Figure 1. Each session consisted of five tasks eliciting distinct behaviors and cognitive workload. Four of them are included in the UDIVA v0.5 dataset, shown in Figure 1 (right), namely:

Talk. Participants were instructed to talk about any subject of their preference for approximately five minutes. This task allows for analysis of common conversational constructs, such as turn-taking, synchrony, empathy, and quality of interaction.

"Animals" game. Participants asked ten yes/no questions about the animal in the card wore on their forehead to guess the animal. Animals were classified into three difficulty levels. This game elicits cognitive processes (e.g., thinking, gaze signaling events).

Lego building. Participants built a Lego together following the instructions leaflet, ranging between four difficulty levels. This task fosters collaboration, cooperation, joint attention, and leader-follower behaviors, among others.

"Ghost blitz" card game. Participants played one card per turn, competing with each other to be the first to select the correct figurine from a set of five figurines placed on the table, based on the content of the played card. This task fosters competitive behavior and allows cognitive processing speed analysis, among others.

The *Talk* task was always administered first as a warm-up, whereas the remaining three were administered in randomized order. The recording of each task starts when the task administrator finishes explaining the task to the participants and stops interacting with them, and finishes when the administrator starts interacting with the participants again to deliver the following task. The real given task (e.g., build a Lego building) may finish minutes before the end of the recording. Once participants finished the real given task, they were free to continue playing or just wait until the task administrator entered the recording room and stopped the recording.

2. https://gdpr.eu/

Figure 1: The UDIVA dataset contains views from six HD tripod-mounted cameras. In UDIVA v0.5, only the frontal views of each participant are provided, FC1 and FC2. Left: Example of the FC1 and FC2 frontal views. Right: Examples of the four tasks (*Talk*, *Lego*, *Animals*, and *Ghost*) from different dyadic sessions.

The UDIVA dataset contains views from six HD third-person cameras and two first-person views. UDIVA v0.5 includes the frontal view of each participant, recorded using two tripod-mounted Revotech i706 cameras at a resolution of 1280×720 pixels and a sampling rate of 25 fps. These correspond to FC1 and FC2 cameras illustrated in Figure 1 (left). The audio was recorded using two Rode smartLav+ lapel microphones, one per participant, at 44100 Hz.

2.2. Statistics

UDIVA v0.5 is composed of approximately 80h of recordings of dyadic interactions between 134 voluntary participants (44.78% female) from 17 to 75 years old (mean=31.95, sd=12.57), mostly Caucasian. Participants come from 22 countries: 74.6% from European (a total of 68.7% from Spain) and 21.6% from Latin American countries. In relation to the maximum level of education, participants had mostly a Master's degree (35.82%), followed by a Bachelor's degree (28.36%). Table 1 shows the descriptive statistics (i.e., mean, standard deviation, and Pearson's correlation) for self-reported personality OCEAN variables for participants of UDIVA v0.5 (see Section 4.1 for a description of Big Five personality traits). Compared to the normative data (Gallardo-Pujol et al., 2021), the UDIVA v0.5 sample presented higher levels of *Open-Mindedness* (O) and lower scores in *Negative Emotionality* (N). A low-to-moderate negative correlation between "N" and *Conscientiousness* (C) and *Agreeableness* (A), and low-to-moderate positive correlation between "A" and "C" and *Extraversion* (E) were observed. Finally, "E" correlated moderately with "O" ($r = .40$) and slightly with "C". These findings are comparable to the pattern found in the literature of intercorrelations between personality traits (Soto and John, 2017; Gallardo-Pujol et al., 2021). Females scored higher than males in all OCEAN values, with statistically significant differences in all traits except in "O". When correlating age and OCEAN scores, medium to low intensity associations were observed, being absolute correlations between 0.04 ("E") and 0.32 ("C").

Trait	Mean	Std. Dev.	O	C	E	A
O	.21	1.07				
C	.12	1.07	.02 [−.15, .19]			
E	−.12	1.01	.40** [.25, .53]	.24** [.07, .39]		
A	−.04	.97	.08 [−.09, .25]	.26** [.10, .42]	.20* [.03, .36]	
N	−.28	1.07	.01 [−.16, .18]	−.22** [−.38, −.05]	−.10 [−.26, .07]	−.22* [−.38, −.05]

Table 1: Descriptive statistics (mean and std. deviation, and Pearson's correlation) for self-reported personality OCEAN values of participants from the UDIVA v0.5 dataset. Values in square brackets indicate the 95% confidence interval for each correlation. $^{*}p < .05$, and $^{**}p < .01$.

Trait	Gender	Age
O	t(132)=0.36; d=0.06	-0.2* [-0.36;-0.03]
C	t(132)=3.04***; d=0.53	0.32*** [0.16;0.47]
E	t(132)=2.28*; d=0.4	0.04 [-0.13;0.21]
A	t(132)=2.74**; d=0.48	0.25** [0.08;0.4]
N	t(132)=3.34***; d=0.58	-0.13 [-0.3;0.04]

Table 2: Differences between genders on OCEAN scores, by means of Student's t-test, and correlations with age. Values in square brackets indicate the 95% confidence interval for each correlation. $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$.

Participants were distributed into 145 dyadic sessions, with a participation average of 2.16 sessions/participant (max. 5 sessions). 44.14% of the interactions occurred among participants that knew each other before their interaction session (i.e., *known* people). Spanish was the interaction language for most of the dyads (73.10%), followed by Catalan (17.25%), and English (9.65%). Regarding the inner state of the participants, the mean pre- and post-fatigue value was 4.35 (± 2.32 for pre-session and ± 2.39 for post-session), whereas descriptive statistics of pre- and post-session mood are shown in Table 3. As can be seen, interaction sessions slightly improved the participants' mood state (i.e., positive states increased, whereas their respective negative counterparts decreased).

	Good	Bad	Happy	Sad	Friendly	Unfriendly	Relaxed	Tense
Pre-session	3.91±0.74	1.83±0.79	3.61±0.78	1.91±0.88	3.86±0.71	1.77±0.81	3.52±0.93	2.34±1.01
Post-session	4.1±0.64	1.67±0.67	3.92±0.72	1.71±0.78	4.04±0.63	1.68±0.73	3.73±0.85	2.07±0.9

Table 3: Descriptive statistics (mean ± standard deviation) of pre- and post-session mood categories of the UDIVA v0.5 dataset.

2.3. Data selection and partition procedure

The interaction sessions included in the UDIVA v0.5 dataset were selected from the complete UDIVA dataset, with the aim of creating subject-independent training, validation, and test splits with similar distribution each in terms of personality traits, age, gender, and relationship among interaction partners. Prior to the data partition process, we first discarded all sessions with participants younger than 16 years old, as such participants filled in personality/temperament questionnaires specific to their age different than the BFI-2 questionnaire (see Palmero et al., 2021), and not all traits measured with such questionnaires have a one-to-one correspondence to the OCEAN personality traits. We also discarded sessions with any major technical issue (e.g., one of the FC views not available or none of the audio signals available for a whole recording).

To ensure that no participants appeared in more than one split, some further sessions needed to be discarded. We followed an iterative procedure to decide which sessions to keep and how to divide them into the different splits. First, we represented the remaining sessions as a graph, where the nodes correspond to participants and the edges correspond to interaction sessions, and assigned weights to sessions based on the number of interactions per participant and the group (i.e., combination of age, gender and binary relationship) they belonged to. That is, initially, sessions with participants that interacted in many sessions and/or their group belonged to a high density one were assigned a lower weight than those sessions with participants that interacted in fewer sessions and/or belonged to less represented groups. Then, we used a greedy approach that iteratively removed and added sessions based on their importance to find split combinations that fulfilled a split ratio of approximately 3:1:1 with respect to the number of participants and sessions. Such approach started removing those sessions with a lower weight. The weights were updated every time a session was removed or added. Once a sample of all possible split candidates was computed, we calculated a set of costs for each candidate based on:

- the difference in per-trait distributions among each split with respect to the sum of splits by means of the p-value obtained from a Kolmogorov-Smirnov significance test (Massey Jr, 1951);
- the differences in Pearson correlation between each personality trait and self-reported gender among each split;
- the differences in Pearson correlation between each personality trait and self-reported age among each split;
- the differences between age, gender, and relationship distributions with respect to a uniform distribution for validation and test splits.

Finally, we selected the combination that minimized the sum of the costs and that maximized the total number of sessions and participants.

The final split contains 116 sessions and 99 participants for training, 18 sessions and 20 participants for validation, and 11 sessions and 15 participants for test. The resulting distribution of OCEAN values is shown in Figure 2. As can be seen, validation and test splits contain a more uniform distribution as intended, trying to correct the characteristic Gaussian-like distribution of personality traits. The age distribution among splits depicted

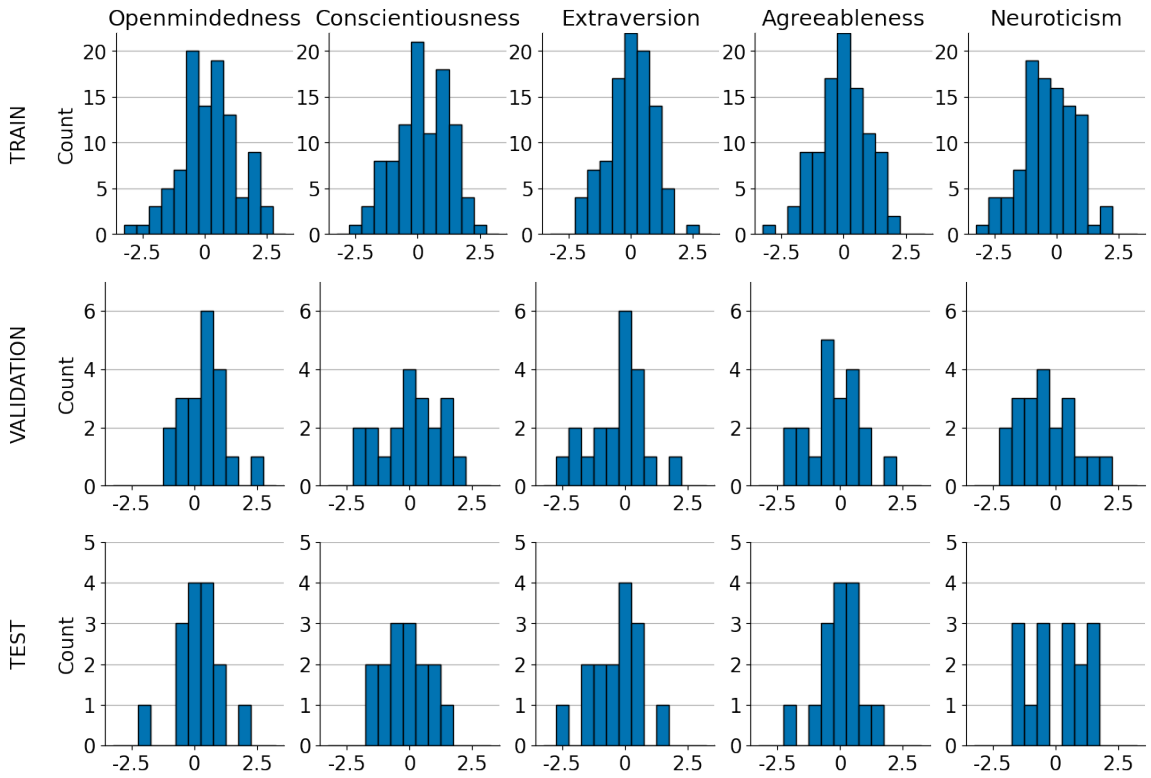

Figure 2: Distribution of the self-reported personality trait (OCEAN) values across train, validation and test splits of the UDIVA v0.5 dataset. The x axis refers to z-scores for each personality trait.

in Figure 3 shows that the test set contains a more restricted but uniform sample than validation. Gender ratios were conserved in all splits. In contrast, relationship ratios are significantly different, having 37.9% of *known* people in training, 61.1% in validation, and 81.8% in test. Pre- and post-session mood and fatigue values in addition to correlations between personality traits and other attributes are shown in Appendix A of the supplementary material. Given that the number of participants in the different splits is low (particularly validation and test), correlations between personality traits and other attributes differed, as expected (Schönbrodt and Perugini, 2013). Indeed, the resulting splits exhibit the difficulty of balancing dyadic data with interdependencies among multiple sessions and a relatively low number of participants. Nonetheless, these splits allow for reliable comparability and benchmarking, especially in the context of the challenge. For other contexts where a higher number of train/test subjects is required, we recommend strategies like leave-one-subject-out or leave-one-dyad-out instead of the provided data splits.

2.4. Content

Here, we describe the data provided with the UDIVA v0.5 dataset for both challenge tracks.

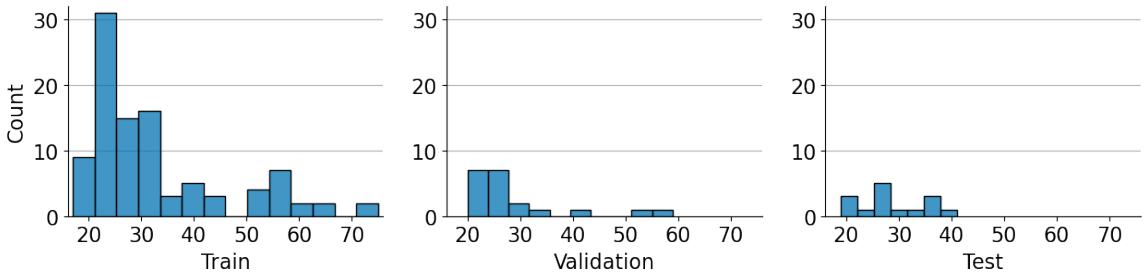

Figure 3: Age distribution across train, validation and test splits of the UDIVA v0.5 dataset.

2.4.1. audiovisual

UDIVA v0.5 dataset includes the two frontal views of the UDIVA dataset, one per participant (FC1 and FC2). Each video is accompanied by a synchronized audio signal from the lapel microphone of the corresponding participant. Both videos are time-synchronized. There is one video per participant, task, and session.

2.4.2. Metadata

Audiovisual data is accompanied by a set of metadata, described next:

Participant. Metadata about each participant, independent of the session. It includes: anonymized participant ID, gender, age, country of origin, maximum level of education, z-score OCEAN personality values, total number of sessions done, session IDs in which the participant has participated in order of occurrence. All information except participant and session IDs, and total number of sessions done, were reported by the participant.

Session. Metadata about the interaction session, as well as participant metadata specific of a given interaction session. It includes: session ID, participant IDs corresponding to each of the camera views, date and time of the recording, difficulty level of *Lego* and *Animals* tasks, language, self-reported relationship among interaction partners (whether they knew each other or not before the session), order of administration of the tasks within the session, pre- and post-session self-reported answers of the mood and fatigue questionnaires per participant, and notes about the session (if any).

Other. Start-end times of the real given task within a task recording (e.g., the time when participants start and end building the requested Lego figurine for the *Lego* task), and start-end times of segments that do not contain audio and/or video because of privacy or technical issues.

2.4.3. Transcripts

Complementing the originally presented UDIVA dataset, literal transcripts of the conversations at utterance level were obtained by a third-party company, and manually reviewed for cleanliness and data protection. As illustrated in Figure 4, a transcript file is composed of: an utterance number, the start and end times of the utterance synchronized to the videos, the ID of the interaction partner (PART.1/2 for the participant visible from FC1/2, respectively), and the transcribed utterance.

No.	Start	End	Duration	Text
44	00:02:34.590	00:02:39.390	00:00:04.800	PART.1: But I was checking and the six, (39) the other one you included... (29)
45	00:02:42.490	00:02:46.590	00:00:04.100	PART.2: So, at the end, (23) it's eleven images you say? (27)
46	00:02:46.690	00:02:47.690	00:00:01.000	PART.1: Yes. (12)
47	00:02:49.690	00:02:50.690	00:00:01.000	PART.2: Really? (15)
48	00:02:50.690	00:02:52.690	00:00:02.000	PART.1: Do you have access here? (32)
49	00:02:52.690	00:02:54.690	00:00:02.000	PART.2: Yes. (12)
50	00:02:57.290	00:03:02.590	00:00:05.300	PART.1: I think you included one (32) that I did not comment (22) but was a good one. (19)

Figure 4: Example of transcript from a short conversational segment included in UDIVA v0.5.

2.4.4. Landmarks and gaze annotations

To further enrich the dataset and provide labels for the Behavior Forecasting track (described in Section 5), we also provided automatically extracted body, face, and hand landmarks, as well as 3D gaze direction vectors. Next, we detail extraction, post-processing, and data cleaning procedures.

Extraction and post-processing. For every frame of each video, the landmarks for the face, body and both hands, as well as the 3D eye gaze direction vector are provided. Their extraction and post-processing are thoroughly described next:

- *Face landmarks.* 68 face fiducials were regressed by the 3DDFA_v2 algorithm presented by Guo et al. (2020)[3]. This method uses a lightweight backbone together with a landmark-regression regularization to achieve state-of-the-art accuracy at very fast speeds. The method also incorporates a short-video-synthesis training strategy, which helps retrieve stabler landmarks for videos. Additionally to the facial landmarks, the face detection confidence provided by FaceBoxes (Zhang et al., 2017) was stored. The face landmarks retrieval was constrained to the most centered face detected within each frame, discarding false detections product of the occasional occlusion caused by the profile view of the other interlocutor. In order to reduce the jitter, the method applied average-smoothing to the landmarks coordinates from frame t with those from frames $t-1$ and $t+1$.

- *Body landmarks.* 24 full-body joints and a detection confidence were retrieved by using the MeTRAbs method (Sárándi et al., 2020)[4], which beat all the tested methods in a preliminary evaluation thanks to performing particularly well with truncated upper bodies. This top-down algorithm detects the body and leverages volumetric heatmaps to extract 2D landmarks in the image space (image coordinates in pixels) along with their corresponding 3D landmarks in the camera coordinate frame. Additionally, detection mistakes were identified and fixed by leveraging a tracker which enforced

3. https://github.com/cleardusk/3DDFA_V2
4. https://github.com/isarandi/metrabs

spatiotemporal continuity (Li et al., 2019)[5]. An extra post-processing step was applied to translate the 3D coordinates to the image space while preserving the depth value. First, we found the 3D similarity transformation T which minimized the least squares problem $TX_{3D->2D} = Y_{2D}$ where $X_{3D->2D} = (x_{3D}, y_{3D})$ (pose camera coordinates), and $Y_{2D} = (x_{2D}, y_{2D})$ (pose image coordinates). The matrix T was defined to have 8 degrees of freedom: 3 for rotation, 3 for translation and 2 for scaling (this parameter was fixed to 1 for z). Then, T was applied to $X_{3D} = (x_{3D}, y_{3D}, z_{3D})$ vector. The resulting $TX_{3D} = Y_{3D}$ vector was stored as the 3D body joints coordinates. This retrieval method did not implement any temporal average smoothing by default. As a result, landmarks had high jitter. To mitigate it, a one-euro filter (Casiez et al., 2012)[6] was applied in the temporal axis (cut off to 0.001 and $\beta = 0.005$).

- *Hand landmarks.* 21 hand landmarks were retrieved with the hand estimator module from FrankMocap (Rong et al., 2021)[7]. This method first detects the hand and then fits a 3D model, which provides 3D hand landmarks. The hand detector leveraged was trained with 100K images featuring a wide range of hand interactions with either objects or themselves (Shan et al., 2020). The hands landmarks estimator was trained with the Ho-3D dataset (Hampali et al., 2020), among others, which contains 60K samples aiming to study the interaction between hands and objects. As a result, the method infers fairly accurate landmarks in the recurrent scenario where hands are interlaced, interacting with objects or mildly occluded (see Table 4 and Figure 5). Similarly to the procedure followed for the body detections, a tracker ensured spatiotemporal smoothness and filled the gaps of hands missed due to rapid movement or severe occlusions (Li et al., 2019). Unfortunately, this scenario is much more challenging and required further processing to increase the quality of the annotations, especially to ensure the consistency of the left-right hand associations. This process is extensively described in Appendix B. The one-euro filter previously described was likewise applied (cut off to 0.001 and a $\beta = 0.02$).

- *3D eye gaze direction vector.* The ETH-XGaze baseline method (Zhang et al., 2020b)[8], based on the ResNet architecture, leveraged the face fiducials previously retrieved to normalize the input images and extract the 3D gaze direction (unit) vectors with respect to the camera coordinate system. This gaze estimation method was trained on ETH-XGaze, a large-scale dataset for gaze estimation purposes, with wide variability in terms of appearance, head poses, gaze directions, and accessories like eyeglasses. The reported error of the method is on par with other state-of-the-art subject-independent remote gaze estimation approaches.

Data cleaning. Additionally, the landmarks from the *Talk* task from the validation and test sets underwent a visual inspection process, since that was the task used to evaluate the methodologies proposed for the Behavior Forecasting track (see Section 5 for a detailed

5. SiamRPN++ implementation from OpenCV 4.5.1 (Python).
6. http://cristal.univ-lille.fr/~casiez/1euro/OneEuroFilter.py
7. https://github.com/facebookresearch/frankmocap
8. https://github.com/xucong-zhang/ETH-XGaze

Figure 5: Face, body, and hands landmarks, and gaze vector displayed over examples of the *Talk*, *Animals*, *Lego* and *Ghost* tasks from 4 sessions (top to bottom rows).

	Validation/Test sets					
	Correct	**Mild**	**Severe**	**Switched**	**VF**	**FBI**
Face	99.3/99.7	0.6/0.2	0.1/0.1	-	-	-
Body	98.1/97.2	1.9/2.7	0.0/0.1	-	-	-
Left hand	88.9/87.0*	7.2/8.0	3.9/5.0	0.4/0.2	2.6/1.6	6.6/7.1
Right hand	90.1/82.3*	5.4/9.7	4.5/8.0	0.4/0.2	2.0/3.4	6.9/7.0

Table 4: Visual annotation process for validation and test sets: prevalence of each label (% of frames). Abbreviations: VF, Visibility Fixed; FBI, Fixed By Interpolation. * Including switched, VF, and FBI annotations.

description). Note that the gaze vector was not visually assessed because it was not evaluated as part of that challenge track. All frames from the 18 validation and 11 test sessions were annotated as follows:

- *Face annotation.* The quality of the face landmarks was assessed and classified for each frame as: 1) *correct*, if all landmarks faithfully matched their anatomical locations, 2) *mild*, if the retrieved face was slightly translated with respect to its anatomical position but its shape and orientation were correct, and 3) *severe*, if either the shape or the orientation was wrong.

- *Body annotation.* Similarly, the quality of the body joints was annotated. Comparatively, we were more permissive regarding the quality thresholds as the body joints estimator yielded noisier predictions in our challenging scenario characterized by truncated bodies. Consequently, body landmarks were classified as *correct* provided that their pose was correct and their individual joints locations matched their anatomical position accounting for certain error. The *mild* label was associated to body landmarks with one inaccurate side that was not caused by occlusion, and the *severe* label to those either with both sides incorrect or with severe single joints mislocations.

- *Hands annotation.* In addition to being wrongly extracted, the hands annotations could present several complementary issues such as left-right mismatch, or being false positive/negatives. In order to distinguish all scenarios, the manual inspection for hands required the annotation of additional labels. First, the quality flag was set to *correct* if the orientation of the annotations was correct and fingers matched their anatomical position allowing for some error. If the fingers did not match either in shape or position but the overall hand orientation was still right, they were labeled as *mild*. On the contrary, if neither the hand orientation nor the fingers were correctly inferred, the hand quality was labeled as *severe*. Additionally, the hand visibility (*visible* or *not visible*, i.e., the hand visibility flag is labeled as *not visible* if the hand is under the table or is behind the body) and cases where the left hand was detected as the right hand or vice versa (hands *switch*) were annotated even when the estimator detected only one hand. Although the post-processing substantially improved the quality of the hand landmarks, wrong or missing hands were still very frequent during

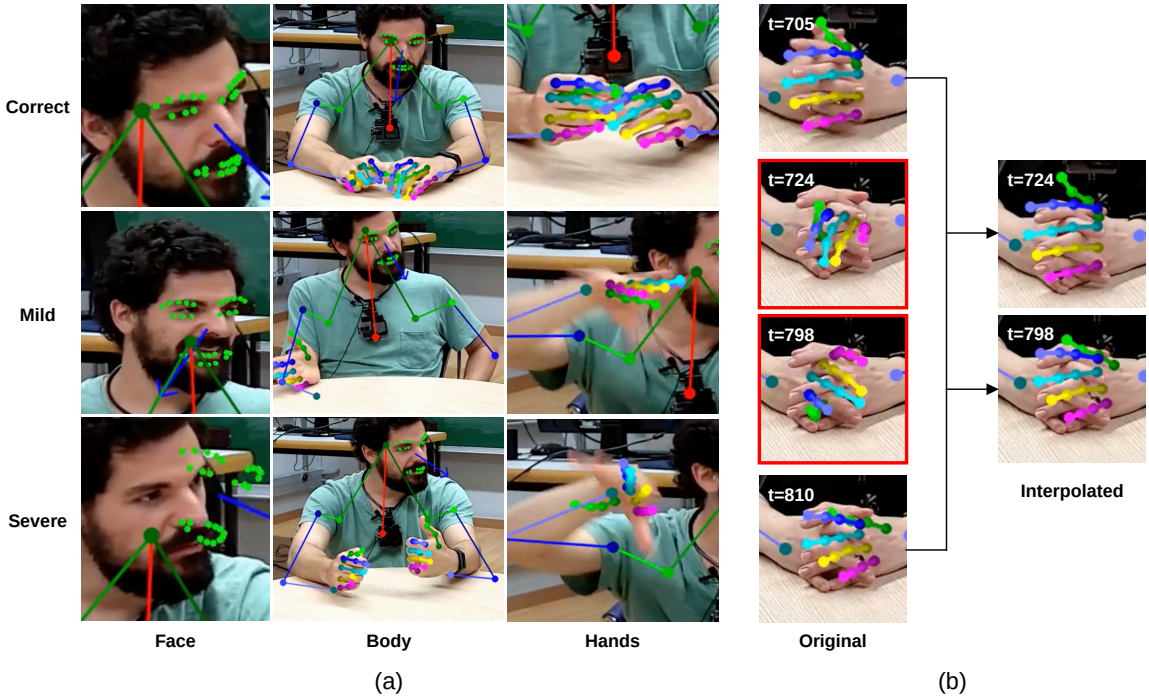

Figure 6: In (a), examples of *correct*, *mild* and *severe* quality labels for face, body and hands landmarks. In (b), a sequence of 104 frames with wrong left-hand landmarks which was fixed by linearly interpolating the landmarks from the last and the first correct extractions before (frame 705) and after (frame 810) the sequence, respectively.

hands interactions or occlusions. In order to maximize the number of correct hand annotations, sequences of consecutive frames $(t_0, ..., t_n)$ with $n \geq 2$ with none or very slow hand movements and correct hand landmarks at t_0 and t_n but with wrong or missing hands in all frames $\{t_i\}_{1 \leq i \leq n-1}$ were identified. From those, segments for which a linear interpolation between the t_0 and t_n landmarks could generate valid hands for the interval (t_0, t_n) were annotated. Such linear interpolation was applied as an extra post-processing step (see Figure 6b).

Figure 6a shows visual examples for the quality labels of each part of the body. Note that, except for the described cases for the hands, the landmarks were not manually corrected. Table 4 summarizes the annotation results for both validation and test sets. The hands linear interpolation proved extremely useful as it fixed up to 7% of the annotations in both validation and test sets (a total of 1714 filled segments for both hands, with an average length of 34.5 ± 78.3 frames). However, the quality of the hands annotations still represents the biggest limitation of the dataset, as the number of *correct* hands annotations varies from 80% to 90% in the validation and test sets.

3. Description of the ChaLearn LAP DYAD'21 challenge

In this section, we describe the organization of the challenge, guidelines to participate, and participation statistics.

3.1. Challenge organization, schedule, and benchmarking platform

The challenge was held in *CodaLab*[9], an open-source web-based platform designed for hosting computational competitions and benchmarking. Challenge participants could participate in either or both of the two tracks, *Track 1: Self-Reported Personality Recognition* and *Track 2: Behavior Forecasting*. To be able to participate, challenge participants were first requested to register to any of the tracks in Codalab, accept the terms and conditions, and fill in and sign the dataset license to request dataset access.

Both tracks were divided into two main phases: *development* and *final evaluation*. Each phase featured a public leaderboard where challenge participants could compare the performance of their proposed solutions against those from other participants and baselines proposed by the organizers given a certain metric to optimize. During the development phase, training data (with ground truth) and validation data (without ground truth) were made available to registered participants, and the submitted predictions over validation data were compared to validation labels. No limit with respect to number of submissions was enforced. Later, during the final evaluation phase, validation ground truth and test data without ground truth were released and the submitted predictions over test were compared to test labels. Test labels were never released during the competition and the participants were limited to 5 submissions during the final evaluation in order to prevent them from improving by trial and error and overfitting to the test data. The final evaluation leaderboard was the one taken into account to determine the potential challenge winners. Both challenge tracks were active from June 1st to September 18th, 2021. The development phase was active throughout the whole competition, while the final evaluation phase ran from September 1st to September 18th.

Once the challenge ended, participants were asked to submit their codes along with detailed instructions on how to run them to reproduce their top-scoring submitted evaluation predictions. Furthermore, they needed to provide a fact sheet detailing their methods. Only those approaches that would outperform the provided baseline could be considered as potential winners, while the remaining approaches could be considered for honorable mentions.

3.2. Provided data and guidelines

Challenge participants were allowed to use all provided data included in the UDIVA v0.5 dataset (unless otherwise stated in each track description below) as well as third-party datasets to train their approaches. All training data was provided unmasked, whereas validation and test data were provided with masked ground truth at the beginning of each corresponding phase. Since the same dataset was provided to all challenge participants regardless of the track in which they participated, and to avoid information leakage, all participants received the data with masked self-reported personality trait values and masked

9. https://codalab.org/

recording snippets to predict from the *Talk* task (i.e., with no audiovisual nor transcript information for those snippets). After each phase, self-reported personality values and information from snippets to predict from the corresponding split were released. For the final evaluation phase, challenge participants were allowed to retrain their models with training and validation data.

3.3. Participation statistics

In total, 67 participants registered for the Self-reported Personality Recognition track, and 44 for the Behavior Forecasting track. From them, we received a total of 20 dataset requests, and a total of 260 submissions, 204 of which were for Track 1 from four unique teams and 56 for Track 2 from three unique teams. Finally, two teams passed the code verification stage for Track 1, and one for Track 2.

4. Track 1: Self-Reported Personality Recognition

In this section, we introduce the task of personality recognition and summarize related work, with a particular focus on recent interlocutor- and context-aware methods. Furthermore, we describe Track 1 of the challenge, the proposed baseline and the top-performing solutions. Finally, we discuss the outcomes and possible future research directions.

4.1. Personality computing

Personality computing (Vinciarelli and Mohammadi, 2014a; Phan and Rauthmann, 2021) is an emergent research field that is attracting increasing attention from different research communities. Personality is widely defined as the manifestation of individual differences in patterns of thought, feeling, and behavior, that remain relatively stable during time (Soto and John, 2017). In the personality computing field, personality is usually characterized by the basic Big Five traits (McCrae and John, 1992) – *Openness to Experience* ("O"), *Conscientiousness* ("C"), *Extraversion* ("E"), *Agreeableness* ("A"), and *Neuroticism* ("N"), often referred to as OCEAN. As described in Vinciarelli and Mohammadi (2014a), personality computing often follows the *Brunswik Lens* (Brunswik, 1956), which is used to describe *externalization* and *attribution* of socially relevant attributes during social interactions. The former is related to personality recognition, which aims to infer self-reported personality traits from observable *distal cues*, i.e., overt behavior. In contrast, the latter is related to personality perception, where the goal is to recognize the *apparent* personality traits of a target person from the perspective of an external observer (Jacques Junior et al., 2019) based on *proximal cues*, i.e., cues that the observer perceives. According to these definitions, we will use the "personality recognition" expression when referring to works that focus on self-reported personality traits (e.g., obtained via self-reported questionnaires), and the "personality perception" expression (or "apparent personality") for referring to works that focus on the personality attributed by external observers. Nevertheless, a method developed for personality recognition could be easily adapted for personality perception and vice versa.

Personality computing has relevant applications in a vast number of scenarios. Personality recognition is increasingly showing its potential to improve well-being and mental health

through personalized interventions (Alexander III et al., 2020). Pedagogical agents (Davis, 2018) which are personalized to maximize learner's attention and learning are another important real-life application. Healthcare delivered by robots is a burgeoning field of research (Esterwood and Robert, 2021), and personalization of robots' personalities seems to have positive impacts on patients' health and social outcomes (Andrist et al., 2015) by means of increasing acceptance of robot's care (Tapus and Matarić, 2008). On a more interactive research field, Hickman et al. (2021) found they were able to profile personality from a job candidate during a job interview interaction with higher accuracy than self-reports, and Song et al. (2021) found that they were able to predict emotional states of an individual through her facial reactions based on cues displayed by the conversational partner. All these benefits rely on the fact that personality is an essential variable in people's interactions with their contexts. The role of personality recognition is to understand the person to the greatest possible degree so as to offer them a personalized version of health, education, or even customer service.

4.2. Related work

Personality recognition has been addressed in the literature using different data modalities, i.e., *still images* (Celli et al., 2014), *image sequences* (Subramanian et al., 2013), *audiovisual* (Batrinca et al., 2011; Lepri et al., 2012a; Fang et al., 2016), *speech and text* (An and Levitan, 2018), and *multimodal* (Anzalone et al., 2017). Apparent or self-reported personality has also been inferred from gaze behavior (Hoppe et al., 2018), mood (Sogancioglu et al., 2021), and even from behavior patterns collected from smartphones (Stachl et al., 2020). Preliminary studies tended to use handcrafted features with standard machine learning techniques (Nguyen et al., 2013; Fang et al., 2016), while more recent works rely on deep learning approaches from handcrafted features (An and Levitan, 2018) or raw data (Mehta et al., 2019).

Most works focus on personality recognition or perception from the individual point of view, even in dyadic or small group conversational scenarios, using only features from the target person. First works that considered interpersonal dependencies for personality computing in face-to-face interaction scenarios used descriptive statistics of overlapping speech segments, short interjections, backchanneling, or interruptions (Pianesi et al., 2008; Valente et al., 2012), or percentages of attention given by the target speaker to other participants and attention received by them (Lepri et al., 2010; Aran and Gatica-Perez, 2013; Subramanian et al., 2013), in addition to other audio/video features of the target speaker. Some works also considered labeled co-occurrent events, such as attention given/received while speaking/not speaking (Lepri et al., 2012b). Okada et al. (2015) proposed to discover frequent co-occurrent events between multiple modalities and people using graph clustering in a small group scenario. In particular, they used utterance segments, speech, gaze, head and body gestures. In a similar scenario, Fang et al. (2016) obtained the highest accuracy when using intrapersonal (speech-, prosodic-, and visual-based), dyadic (speech-based, such as speaking interruptions and backchanneling), and speech-based one-versus-all features.

Most of the aforementioned methods rely on handcrafted interpersonal features. However, few methods propose interlocutor- or context-aware architectures. The work of Su et al. (2016) was one of the first for dyadic conversations, but focusing on personality

perception. They relied on a recurrent network to model the relationship between the linguistic features of each speaking turn and personality, and on coupled Hidden Markov Models to then model the long-term turn-taking temporal evolution and cross-speaker contextual information to infer the personality of both individuals. Later, Lin and Lee (2018) proposed an interlocutor-modulated recurrent attention model with turn-based acoustic features, which models the vocal self and interactive behaviors of a target speaker during small group interactions. Zhang et al. (2020a) predicted self-reported personality and performance labels by correlation analysis of co-occurrent intrapersonal and interpersonal key action events, extracted from head and hand pose, gaze, and motion intensity features. Regarding context, Principi et al. (2019) were one of the first to consider person metadata (e.g., gender, age, ethnicity, and perceived attractiveness) with audiovisual data. However, their goal was to better approximate the crowd biases for apparent personality recognition in one-person videos. More recently, Palmero et al. (2021) proposed a Transformer-based method for self-reported personality recognition in dyadic scenarios, which uses audiovisual data and different sources of context from both interlocutors and the tasks performed to regress a target person's personality traits. In their work, the individual's personality was regressed from 3-second snippets, which may not be enough to properly model long-term interactions. Multimodal fusion was done by simply concatenating the information from the video and audio modalities. To address these limitations, Curto et al. (2021) exploited variable time windows, allowing the capture of long-term interdependencies (e.g., longer video snippets up to 30 seconds), and cross-subject layer, which allows the network to explicitly model interactions among interaction partners in a dyadic scenario through attentional operations. In their work, the behavior from both individuals is explicitly modeled simultaneously through a two-stream cross-attentional Transformer, to eventually predict their self-reported personalities jointly. Very recently, Shao et al. (2021) proposed to infer an individual's personality by modeling their cognitive processes. More concretely, the approach first learns a person-specific convolutional network that predicts the target's facial reactions to the other speaker's audiovisual cues (facial landmarks and Mel-frequency cepstral coefficients). Then, personality is inferred from the graph representation of the target's person-specific processor. Since such processor is optimized on all available data of the target, it can use the information from entire videos at once to infer personality, not only short video snippets.

4.3. Description of the competition track

The goal of the Self-Reported Personality Recognition track was to regress the self-reported per-subject OCEAN personality traits, having at disposal all available data in the UDIVA v0.5 dataset. Challenge participants had to outperform the provided baseline results in terms of the average Mean Squared Error (MSE) between the ground truth and the predicted scores, taking into account the individual score of the different traits.

4.4. Baseline approach

As the baseline for this track, we proposed a multimodal attention-based model that receives different sources of information and context from both participants in the dyadic interaction

to regress the target person's personality traits. An overview of the method is described below, while the complete details can be found in Palmero et al. (2021).

The architecture is a re-purposed Video Action Transformer Network (Girdhar et al., 2019), the input of which consists of video, audio, and metadata information. The time-synchronized full-length videos corresponding to each interlocutor were split into 32-frame length non-overlapping chunks, capturing approximately 2.5 seconds of information. They are fed to a pretrained R(2+1)D (Tran et al., 2018) network, and associated features are extracted from the third convolutional residual stack. The face of the target person is extracted from the full-frame chunks using a pretrained MobileNet-SSD (Howard et al., 2017) and processed similarly. Time-synchronized audio features were extracted from the raw audio signal of each participant's lapel microphone using a pretrained VGGish (Hershey et al., 2017) model. Finally, provided characteristics of both participants were used along with session and dyadic metadata. More concretely, as participant metadata we used: age, gender, cultural background (Mensah and Chen, 2013), the number of times that participant had been recorded so far, and pre-session mood and fatigue values. As session metadata, we used the order of the task within a session and the task difficulty, using 0 for tasks with no associated difficulty level. As dyadic metadata, we considered the participants' relationship.

A fundamental characteristic of the Action Transformer network is the selection of the query, keys and values. In the case of the proposed baseline, the query incorporated the face and the metadata of the target person. We considered two types of keys and values: local and extended. Local keys/values included audiovisual embeddings from the target person, while the extended counterparts included audiovisual and metadata embeddings from the other interlocutor. The local and extended key and value embeddings together with the query are processed independently in two different units of the transformer layer. They provide two updated queries that are concatenated and linearly projected to produce the final context-updated query. The baseline architecture uses three such transformer layers.

Since the model receives chunks of the original videos as input, the predicted personality traits are obtained at chunk level. We computed the median between all the chunk scores corresponding to each person to obtain the final participant-wise personality values.

4.5. Top-performing solutions

The scores of the challenge participants for each personality trait and the average score are reported in Table 5. As can be seen, the SMART-SAIR team outperformed the baseline for all the personality traits and was declared the winner of the personality track. The remaining participants were not able to improve the baseline's average score. However, the FGM Utrecht team was awarded an honorable mention for achieving the highest score for "C". The table also includes the error obtained by an additional *mean prediction* baseline, which uses the mean of the per-trait ground truth personality labels of the training set as the prediction for the individuals on the test set. We observe that the winner and honorable mention methods outperform this mean baseline on average and for all traits except for "O", for which this baseline is only outperformed by the winner team.

Team	Avg. Score ↑	O	C	E	A	N
Mean prediction	0.889	0.725	0.877	0.991	0.673	1.179
SMART-SAIR	0.7691	0.7112 (1)	0.7230 (2)	0.8665 (1)	0.5482 (1)	0.9966 (1)
Baseline	0.8179	0.7443 (2)	0.7935 (3)	0.8864 (2)	0.6531 (2)	1.0121 (2)
FGM Utrecht	0.8251	0.7522 (3)	0.6874 (1)	0.9174 (3)	0.6707 (3)	1.0978 (3)
STARS Inria	1.0580	0.8386 (4)	0.9762 (4)	1.3586 (4)	0.8644 (4)	1.2520 (4)

Table 5: Codalab leaderboard (*final evaluation* stage) of Track 1: Self-Reported Personality Recognition. Per-trait scores are reported by means of Mean Squared Error (MSE). Values within parentheses denote the rank position. The Average MSE per method is also reported. Methods that passed the code verification stage and described in this paper are highlighted in bold. ↑ indicates the lower the score, the better. *Mean prediction* refers to the performance of a system that returns the average per-trait personality ground truth labels of the training set as the predicted personality.

Detailed information about the winner and honorable mention methods can be found in their respective fact sheets: SMART-SAIR[10] and FGM Utrecht[11]. Both solutions are briefly described in the following subsections.

4.5.1. Winning solution: SMART-SAIR

Motivated by the findings that significant differences in personality traits between men and women exist (Weisberg et al., 2011), the SMART-SAIR team decided to cluster the dataset participants into two profiles based on their gender, and developed a multimodal model for each profile using Neural Architecture Search (NAS, Jin et al. 2019). The proposed pipeline is illustrated in Figure 7. The features used to train each model come from different modalities: visual and textual (i.e., transcripts).

1. **Visual features** consisted of facial and body pose landmarks obtained from the annotations provided with the dataset. For each video frame, 68 3D facial landmarks are used, obtaining a flattened 204-dimensional array. For the body, 24 3D landmarks are used, obtaining a flattened 72-dimensional array. The mean and standard deviation of each facial and body pose landmark point over all the frames in a 1-minute video clip are computed, resulting in 2 feature vectors of sizes 408 and 144, respectively. Then, the face and body landmark statistics are concatenated, resulting in a 552-dimensional feature vector for each 1-minute video clip.

2. **Text-based features** are obtained through the analysis of the transcripts, based on each *talk-turn* (i.e., utterance) duration, content, and sentiment.

    - *Duration:* the duration of interaction for a person in a single minute was analyzed to generate a 5-dimensional feature set consisting of the minimum and maximum

10. https://chalearnlap.cvc.uab.cat/media/results/None/Track-1_top-1_ICCV_Learning_Personalised_Models.pdf
11. https://chalearnlap.cvc.uab.cat/media/results/None/Track_1-Honorable_Mention_Fact_sheet__Challenge.pdf

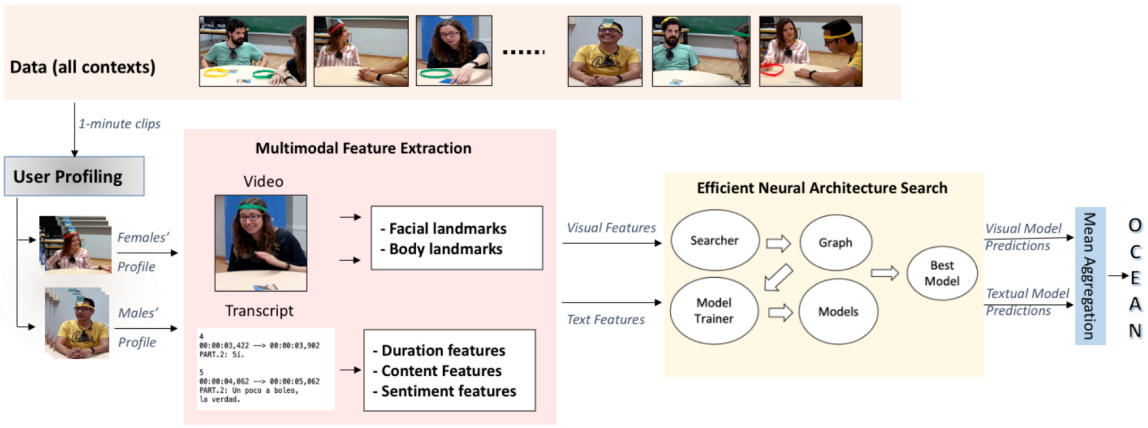

Figure 7: SMART-SAIR workflow approach[10] for self-reported personality recognition.

turn duration, average turn duration, standard deviation of turn duration, and total duration of turns.

- *Content:* the features extracted by analyzing the number of turns and the content of the dialog is represented as a 5-dimensional vector produced by the following features: the percentage of turns for a particular person out of the total number of turns in a single minute, average words per turn, longest turn, total number of words, and standard deviation of words per turn.
- *Sentiment:* Each of the 1-minute transcripts was analyzed to generate 10 sentiment-based features using a Spanish sentiment recognition library[12], as it is the predominant language across all sessions. The generated sentiment values ranged in between 0 to 1, where 0 corresponds to fully negative and 1 corresponds to fully positive sentiment. Then, different features were obtained from that, i.e., the most negative/positive turn values, average sentiment per person, sentiment range, and overall sentiment.

Having the model automatically designed using NAS and trained for the different profiles based on gender, the SMART-SAIR team applied decision-based fusion to predict the personality of each individual. That is, the scores obtained per minute were averaged over all the sessions. Then, the resulting values were aggregated across the different modalities using the average predictions of both modalities.

According to the SMART-SAIR team[10], the proposed approach has two main advantages. Firstly, it combines multiple modalities to predict the personality of an individual. Secondly, the system is scalable and can adapt itself to changing trends in the data as the neural architecture search-based approach enables the generation of a deep learning model depending on the user profile. Finally, they also observed that visual and textual features performed almost equally well when used individually, and that the best results were obtained by combining both the visual and textual features.

12. https://pypi.org/project/sentiment-analysis-spanish/

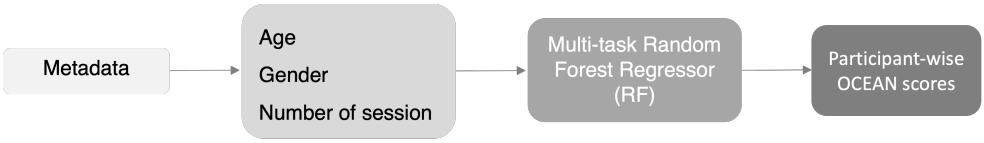

Figure 8: Final solution of the Utrecht team[11] for self-reported personality recognition.

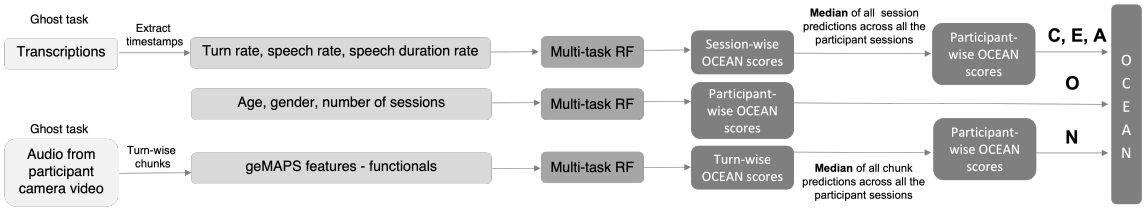

Figure 9: Alternative solution of the Utrecht team[11], used during *development* phase.

4.5.2. Honorable mention: FGM UTRECHT

The FGM Utrecht team analyzed the correlation between metadata and the self-assessment personality trait scores, and proposed to use a Random Forest (RF) regressor trained solely on metadata features (i.e., age, gender and number of sessions). The proposed pipeline is illustrated in Figure 8.

Alternatively, they also proposed to combine linguistic, audio and metadata features in a multitask fashion, based on RF regressors and late fusion, illustrated in Figure 9. In this alternative solution, they trained different regressors from different data modalities, based on a preliminary analysis using the *development* data (i.e., evaluated on the validation split). Interestingly, they observed that different traits were better predicted from different modalities. For instance, the "O" trait was better predicted from metadata information; the "C", "E" and "A" traits were better predicted using text-based features; and the "N" trait was better predicted from audio features. Furthermore, they also found that the best performing models for both linguistic and audio features were obtained in the *Ghost* task. However, such methodology was not able to outperform their simple RF regressor trained solely on metadata for the *final evaluation* phase (i.e., evaluated on the test split), despite doing so for the *development* phase. Detailed information about such alternative solution can be found in their fact sheet[11].

According to the FGM Utrecht team[11], obtained results validate to some extent the findings that a correlation between personality, gender, and age exists (Marsh et al., 2012). They also suggested that the number of sessions may be correlated with the *Openness* trait, considering the time investment and social energy required to participate in these sessions. In their conclusions, they commented that a simple RF regressor based on metadata features only should not be enough to outperform a method based on linguistic, audio-visual and metadata features like the alternative model they evaluated in such a complex task like personality recognition, indicating that the UDIVA v0.5 dataset may include unwanted bias. Finally, they commented that the possible sources of bias could be associated with the small number of participants in the dataset, the similarities in their backgrounds, and

education level, which might not be representative enough for building a strong model capable of generalizing in a more diverse population.

4.6. Discussion

Several lessons can be drawn from the self-reported personality recognition track. Both winning and honorable mention solutions suggested that a correlation between personality and gender exist, which is supported by the literature on the field (Weisberg et al., 2011; Marsh et al., 2012). Although previous work on personality computing identified the existence of distinct types of gender-bias on related datasets (Principi et al., 2019; Escalante et al., 2020; Jacques Junior et al., 2021), most methods found in the personality computing literature are proposing to advance the research on the topic without explicitly taking such correlation into account. Although personality traits are often regarded as relatively stable over time (Cuperman and Ickes, 2009), they actually follow specific patterns of change that have been already described (e.g.,"C" increases over lifetime, Roberts et al. 2006). Personality values denote behavior tendencies averaged across many situations; therefore, self-reported personality predictions from short temporal segments can be noisy and later aggregation of those should be carefully designed. In fact, within-person variability is greater than between-person variability (Fleeson and Jayawickreme, 2015), even during face-to-face interactions (Gundogdu et al., 2017), thus point estimates of personality traits should be taken with caution. The use of longer time windows capable of capturing long-term interdependencies (Curto et al., 2021) or the learning of person-specific networks leveraging complete interactions (Shao et al., 2021) are potential alternatives to address this problem. Nevertheless, the modeling of context-aware long-term interdependencies is complex, particularly in the case of multimodal scenarios.

Results from challenge participants have also shown that vocal behavior (both speech features and transcripts) provides rich cues for personality recognition, which is aligned with previous work on personality perception showing that audiovisual models can benefit from the inclusion of language data (Güçlütürk et al., 2018). Furthermore, both teams have shown the potential of using lower-dimensional input representations for their personality models, in contrast to the baseline approach that uses raw data. Results have also revealed a recurrent and consistent finding in personality computing (Jacques Junior et al., 2019) that there is no single model that works best for all possible settings and personality traits, reinforced in Section 4.5.2 by the FGM Utrecht's solution. This suggests that different traits can be better modeled and recognized by distinct feature representations and methodologies.

Fairness in machine learning can be defined in different ways (Kusner et al., 2017; Friedler et al., 2019). In general, most definitions propose to fight against algorithmic biases in order to ensure that the decisions of any machine learning-based method do not reflect discriminatory behavior toward certain groups or populations (Mehrabi et al., 2021). In the context of decision-making, and according to Mehrabi et al. (2021), fairness is the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics. Thus, an unfair algorithm is one the decisions of which are skewed toward a particular group of people. Therefore, if we want the next generation of intelligent systems for personality computing to be fair with respect to different contexts, setups and demographics, we need to address the fairness problem in some way, either through the design

and development of new datasets and annotation protocols or through new methodologies capable of mitigating different types of bias. For instance, the winning approach of the personality recognition track proposed to use as visual information facial and body landmarks. Facial landmarks detectors may have their own biases, depending on how they are trained. They also have a potential power to reveal racial information based on face morphology and geometric features (Bülthoff et al., 2021), which could be considered a possible source of bias. However, facial landmarks can also be used to remove some appearance information that could be another and probably even stronger possible source of bias due to stereotyping, such as skin tone, hair style or face attractiveness. A future research question to be addressed on this domain could be: *is it possible to build generic features free of bias with respect to demographics attributes (e.g., gender, age, race) for different kinds of modalities (e.g., audio, visual, text)?* Completely eliminating demographic bias is extremely difficult, if not impossible, due to unknown factors or correlated features. However, moving toward fair machine learning is crucial to advance on the use of artificial intelligence for the good of society. Automatically identifying whether the outcomes of a particular model are biased toward different demographic groups of individuals is another highly relevant future research direction on this domain, as well as the automatic identification of unknown biases.

In this regard, UDIVA v0.5 is an unbalanced dataset. For instance, although the dataset is relatively balanced with respect to the gender attribute, it is extremely unbalanced regarding age, country of origin, or spoken language. Besides that, personality recognition is treated as a regression task where not all the range of values are equally represented, usually following a normal distribution (McCrae and John, 1992). As mentioned in Section 2.2 and raised by the FGM Utrecht team, the dataset is slightly biased towards people with high levels of the *Openness* trait due to the recruitment procedure. These and other possible sources of bias make the problem harder, but also stimulate research on bias identification and mitigation methods and related areas, in addition to the multimodal personality computing problem by itself. Another limitation of the UDIVA and UVIDA v0.5 dataset is the number of subjects, i.e., 147 and 134, respectively. While being one of the face-to-face interaction datasets with such a high number of participants (Palmero et al., 2021), correlations are known to achieve their point of stability around 160 subjects for typical scenarios in personality psychology, achieving full stability at 250 (Schönbrodt and Perugini, 2013). From a psychology standpoint, this means that our findings reveal trends that will not radically change, but can vary in terms of size, not in terms of the direction of correlation. Thus, our findings are reliable. However, from a machine learning point of view, where we aim to infer personality from overt behavior, the need for further data that covers a broader spectrum of behavior combinations significantly increases. Consequently, our study can motivate the research and design of novel datasets and annotation protocols, in particular large-scale datasets for unsupervised or self-supervised pre-training for shared, social-related downstream tasks.

5. Track 2: Behavior Forecasting

Similarly to Section 4, in this section, we first introduce the task of non-verbal social behavior forecasting, potential applications, and technical related work particularly focusing on landmark-based approaches. Then, we describe the challenge baseline and top-performing

solutions of Track 2. The section ends with a discussion of the outcomes of this track and suggestions for future research.

5.1. Social behavior forecasting

Humans have the innate ability to anticipate actions continuously, given a set of previous observations. Such anticipation is crucial in many time-sensitive contexts. For example, Sexton et al. (2018) showed that around 30% of requests during surgery were anticipated by the assistants, which translated to significant operating time reductions. The influence of anticipation is not limited to such specific environments though. In fact, similar efficiency improvements have been observed in other robot-assisted tasks like object grasping (Saponaro et al., 2013), or executing actions towards a common goal (Dominey et al., 2008; Huang and Mutlu, 2016). Behavior forecasting has also applications in assistive technology (e.g., anticipating falls), collaborative and autonomous robots or autonomous vehicles (Chaabane et al., 2020). For instance, Gujjar and Vaughan (2019) achieve to decode the behavior of pedestrians based on previously recorded videos to inform the decisions taken by autonomous vehicles. Another application works with visually impaired populations, in which behavior forecasting has helped predict the future trajectory of people in egocentric settings (Qiu et al., 2021).

In a more interactive and social context, Duarte et al. (2018) showed the benefits of modeling non-verbal signals exchanged during interactions to make an embodied agent act in an understandable and socially aware way. In this direction, the research community has explored the forecasting possibilities for many human behavior representations in social contexts such as feedback responses, disengagement, or turn-taking (Barquero et al., 2022a). For instance, in dyadic interactions, behavior forecasting has been applied in scenarios ranging from the prediction of non-verbal behavior allowing more human-like interactions with artificial interfaces or avatars (Ahuja et al., 2019), to the prediction of periods of agitation in people with dementia (Alam et al., 2019).

5.2. Related work

Low-level representations, such as body or face landmarks, are commonly used as input for single human behavior forecasting (Sofianos et al., 2021; Parsaeifard et al., 2021), although few works leverage them in social contexts (Katircioglu et al., 2021; Guo et al., 2021). The stochastic nature of future prediction makes behavior forecasting an extremely challenging problem. For instance, in the much more generic problem of future video prediction, the future uncertainty leads to blurred predicted sequences of frames (Oprea et al., 2020). While unavoidable, the ambiguity can be reduced by narrowing the space of predictions and increasing the contextual information available. Regarding the prediction space, a common path is to forecast low-dimensional data such as landmarks or action units. As a matter of fact, in the past years, human body poses and facial expressions have been leveraged to recognize behavior (Nadeem et al., 2020), infer intent (Kim et al., 2020) or detect behavioral and emotional disorders (Vyas et al., 2019). Therefore, forecasting landmark representations can be seen as an intermediate step towards the forecasting of any visual behavioral cue or signal (Stergiou and Poppe, 2019).

Most behavior forecasting works focus on single-human scenarios. Existing methodologies have traditionally followed two main directions. Some of them embrace the aforementioned future uncertainty and exploit it by predicting multiple futures, which are usually referred to as stochastic methods. Most common approaches are conditional variational autoencoders (Yuan and Kitani, 2020; Kania et al., 2021; Zhang et al., 2021b) and conditional generative adversarial networks (Barsoum et al., 2018; Fushishita et al., 2019). The main problem derived from these methods resides in their evaluation, as the ground truth only includes one sample of all hypothetical futures. Instead, other works simplify it by considering the future to be deterministic. Compared to the stochastic methods, the results of deterministic methods can be easily evaluated by directly comparing them to the ground truth. Most of them add constraints or provide contextual information in order to narrow the future space and therefore reduce its stochasticity. Usual methodologies include encoder-decoder recurrent approaches (Fragkiadaki et al., 2015; Chung et al., 2016; Martinez et al., 2017; Ghosh et al., 2017; Chiu et al., 2019) and Transformer-like architectures (Mao et al., 2020; Cai et al., 2020). Other works underline the importance of incorporating scene knowledge to the human behavior model, including the objects position and trajectory (Shu et al., 2016; Corona et al., 2020b; Adeli et al., 2021), or visual spatiotemporal features of the scene (Adeli et al., 2020, 2021).

In contrast, few works attempt to predict behavior on conversations between two or more interlocutors, and none tackles them from a stochastic point of view. When a group of people is interacting, the positions and orientations of the individuals, their body and hand gestures, gaze, and facial expressions become extremely relevant for behavior forecasting. In the context of face, most works aim at predicting lower-dimensional representations of the face such as action units (Chu et al., 2018; Chen et al., 2019; Ueno et al., 2020; Woo et al., 2021), facial landmarks (Huang and Khan, 2017), or explicitly learned representations (Feng et al., 2017). For body pose forecasting, methodologies can be grouped by how the landmarks are handled: in the image or in the frequency space. The former refers to classical approaches, with encoder-decoder architectures based on recurrent neural networks such as Long Short-Term Memory (LSTM, Ahuja et al. 2019; Hua et al. 2019; Honda et al. 2020) or gated recurrent units (Adeli et al., 2020) as the gold standard, sometimes assisted by graph neural networks (Corona et al., 2020b; Adeli et al., 2020, 2021). Instead, the latter approaches apply the Discrete Cosine Transform (Ahmed et al., 1974) to each body joint's temporal sequence to process them in the frequency space. This favors the intra- and inter-person attention with Transformer-like architectures (Wang et al., 2021b; Guo et al., 2021; Katircioglu et al., 2021; Wang et al., 2021a) and reportedly avoids generating freezing motion. The only available works on hand-gestures forecasting tackle other problems in non-social environments such as grasping object affordance (Corona et al., 2020a) or early gesture recognition (Lee et al., 2018). Few works try to leverage multimodal information like verbal content (Chu et al., 2018; Hua et al., 2019; Ueno et al., 2020), prosodic features (Ahuja et al., 2019; Ueno et al., 2020; Woo et al., 2021), or visual features (Adeli et al., 2020, 2021). Naturally, multimodal data needs to be exploited in a specific way in order to fully profit from it.

5.3. Description of the competition track

The goal of the Behavior Forecasting track consisted in using information from the past in order to predict the position of the face, body and hand landmarks in the future within a given time window, in the form of 2D coordinates in the image space. The length of this *prediction window* was set to two seconds (50 frames), which allows us to evaluate both short- $(0 - 400ms)$ and long-range behaviors $(> 400ms)$ (Tang et al., 2018; Oprea et al., 2020; Kania et al., 2021). In order to provide a reasonable conversational context, the information from the immediate previous frames of the prediction window is needed (*observation window*). Each pair of observation and prediction windows will be hereafter referred to as *segment*. These segments conceptually include information of both individuals. Although we enforced all validation and test segments to have observation windows of at least four seconds (100 frames), there were no specific requirements regarding the time window that challenge participants had to use to make their predictions.

The *Talk* task was selected to evaluate this track due to its highly interactive nature, which could provide the most valuable insights about the underlying mechanisms of human-human interaction. Prediction window snippets were defined for this task (explained in Section 5.3.2) for validation and test data. In addition to the nature of the task, the landmarks extraction process was less challenging in this scenario due to the lack of occlusions and interactions with objects. Nevertheless, challenge participants had to deal with a relatively high amount of noise in the labels available: the evident jitter for the body limbs, the often wrong wrist joints from the body landmarks, and the missing or wrong extractions of hands landmarks. This encouraged the exploration of methods especially suited for such weakly supervision and methods that explicitly leveraged noise mitigation mechanisms.

In contrast to the Personality Recognition track, where challenge participants could use all available data to infer per-subject personality traits, participants of the Behavior Forecasting track were only allowed to use information from or obtained in the past to predict the behaviors occurring in a given prediction window. That discards post-session fatigue and mood values provided as metadata, as well as further information coming from other interaction sessions if such sessions took place after the session to which a given prediction window belongs. Consequently, as the *Talk* task (i.e., the one used for evaluation) was always administered first during a session, the remaining tasks of that session counted as future information as well. Therefore, it was not allowed to use information coming from those future tasks either. In practice, this implies that the first predictions for a given subject may be worse than future ones, as more information about the subject is being provided and used sequentially. In the code verification stage, we carefully checked that the participants did not use information from the future to predict the given prediction windows.

5.3.1. Pose representation

In order to reduce the complexity of our problem, the evaluation only considered a subset of the most relevant landmarks for human behavior analysis. For face, only 28 of the 68 original landmarks were considered: five landmarks per eyebrow, three landmarks per eye (four inner points averaged as the eye center, and both end points), five landmarks per lip (computed by averaging the top and bottom edges of each lip), and the two mouth extremes.

From body, we only considered the 10 upper-body joints (both wrists, elbows, shoulders and chest landmarks, and torso and neck). We kept all the landmarks of the fingers (20 in total, four for each finger) and discarded the wrist one, which was found to be much noisier than the one from the body pose. Several examples of the selected landmarks are shown in Figure 6.

5.3.2. Segments generation

In order to fairly evaluate dyad-driven behaviors, the landmarks from the evaluation segments had to be as accurate as possible for both session participants. Ideally, the evaluation could have been constrained to the 150-frames-long sequences (i.e., segments) with *correct* face, body, and hands landmarks for both participants. However, this restrictive scenario implied an upper bound for the amount of non-overlapping segments of 288 for the validation set and 97 for the test set (28.2% and 16.2% of total frames used, respectively). In order to keep a good trade-off between landmarks accuracy and number of segments fulfilling the constraints, the *correctness* condition was only required for face, body, and at least one hand (if visible) for both participants. This increased the previous upper bound up to 658 and 300 non-overlapping segments for validation and test sets, respectively.

The set of evaluation segments had to be as diverse as possible to represent the widest spectrum of human behavior possible within the dataset. Unfortunately, such behavioral diversity is difficult to model, especially without any high-level behavioral labels. We simplified it by assuming that such variety could be preserved by enforcing the diversity of movement velocities (magnitude and angle) and hands visibility and correctness. Therefore, for each candidate segment, the angles and speeds of the 28 face landmarks, six body arms landmarks and 20x2 hand landmarks were computed. For each participant, two 2D histograms were generated for each landmark: one for the observation window and one for the prediction window. The histograms were computed with three bins for speed ($[0, 3)$, $[3, 6)$ and ≥ 6 px/frame for face and body, and $[0, 4)$, $[4, 8)$ and ≥ 8 px/frame for hands) and 11 bins for angles (equally split from $-\pi$ to π) and normalized. The raveled histograms were concatenated along with the percentage of correct/visible hands (eight values, two for each hand of both participants), yielding a feature vector of size 9784 per candidate segment (4892 for each participant). These vectors from the validation set were grouped into 16 clusters (chosen with the elbow method) with the K-means algorithm (Hartigan, 1975). To ensure a uniform distribution of segments across clusters, the final segments were greedily sampled by selecting all non-overlapping candidate segments starting from the least populated cluster. This strategy yielded 598 validation segments (37.4 ± 10.4 segments per cluster). The same clusters were used to classify and sample the test candidate segments, resulting in 278 test segments (18.5 ± 13.6 segments per cluster). The distribution plots of both candidate and final segments for the validation and the test sets are shown in Appendix C.

Masking of prediction window snippets for validation and test data was done at frame level for video and audio, using the timestamps of the start and end frames of the snippet for the latter. Utterances with start and end times between the timestamp of the first frame and the timestamp of the last frame of the snippet were removed. Additionally, utterances that started before the prediction window and continued into such window, and/or those that started during the prediction window and finished later, were manually masked by

modifying the start/end utterance timestamp with the start/end prediction window timestamp, and replacing the utterance content that took place during the prediction window with a [MASKED] note.

5.3.3. Metrics

The objectives of the Behavior Forecasting track are accurately predicting face, body, and hands motions at the same time. A suitable metric for such goals should: 1) encourage realistic future predictions, and 2) avoid penalizing outlier predictions dramatically. The first constraint prevents us from using any distance metric (e.g., standard mean per joint position error or MPJPE, c.f. Ionescu et al. 2013), as any corrupt prediction would penalize the final score too heavily. Instead, upper-bounded metrics could be considered. The second constraint is quite challenging; while realism can be easily assessed from a qualitative point of view, providing a quantitative score is not trivial. Instead, we propose a set of broadly adopted metrics in pose estimation that already implicitly imply such realism. Note that, in order to account for the variability in the dataset participants' anatomy, the errors in the metrics are always computed up to an anatomical measure specific to each participant.

Face. Area Under the Curve (AUC) of the Cumulative Error Distribution plot (CED). It was computed up to the 25% of the inter-pupil distance. Such distance was defined by averaging the distance between the 3D landmarks of both eye centers in all participant frames. Although the most common upper bound for the CED in the literature of face landmarks estimation was 10%, we relaxed it due to the superior difficulty of the forecasting problem (Valstar et al., 2010). As a result, the face metric is sensitive to accuracy improvements in a wider error spectrum while still significantly rewarding the most accurate predictions. Furthermore, in order to avoid rewarding face predictions with sparse error distributions, we computed the CED with the maximum error across landmarks, instead of the average. Therefore, our metric only rewards faces that preserve a correct facial shape up to some extent. Note that only frames with *correct* face landmarks were considered (M_F). 100 bins were used for the AUC calculation.

$$F := AUC^{M_F}_{CED(0:0.25)}$$

Body. Average of the Percentage of Correct Keypoints (PCK) up to 50% of the head size (Andriluka et al., 2014) for all the body joints. While high values for this metric do not implicitly imply body realism, we do not consider this a problem given the static nature of the body in seated conversational scenarios. Only frames with *correct* body pose were considered (M_B). Wrists joints were only evaluated when the hand was annotated as *visible*. In order to contemplate the higher levels of noise present in the body annotations, only ten bins were considered for the PCK scores computation.

$$B := \frac{1}{P}\sum_{i=0}^{P} AUC^{M_B}_{PCK_i(0:0.5)}, \quad \text{where} \quad P := \#\text{body-joints considered}$$

Hands. AUC of the Success Rate (SR) (Yuan et al. 2018) plot computed up to the 50% of the palm size. The palm size was defined as the distance between the knuckles of the index and the little fingers. Similarly to the face scenario, we considered the maximum

Team	Average Rank ↑	Face	Body	Hands
Baseline	1.00	0.3458 (1)	0.8897 (1)	0.5392 (1)
rays2pix	1.67	0.3458 (1)	0.8824 (2)	0.5177 (2)
SAIR KCL	2.67	0.2049 (2)	0.8507 (3)	0.3160 (3)

Table 6: Codalab leaderboard (*final evaluation* stage) of Track 2: Behavior Forecasting. Methods that passed the code verification stage are highlighted in bold. ↑ indicates the lower the score, the better.

error across all landmarks. The final hands score was the average of this value for both hands. Only frames with *correct* hand landmarks were considered (M_L and M_R). 100 bins were considered for the AUC calculation.

$$H = \frac{H^L + H^R}{2} = \frac{(AUC^{M_L}_{SR_L(0:0.5)} + AUC^{M_R}_{SR_R(0:0.5)})}{2}$$

Challenge participants were ranked according to the three metrics independently. The team with the highest average position in the three rankings which surpassed the baseline position would be declared the winner.

5.4. Baseline approach

Our baseline for the Behavior Forecasting track consists in propagating the landmarks from the last observed frame into the future, as if the person remained static once the observation window finished (*zero-velocity* baseline). While it may seem counterintuitive, the zero-velocity baseline has proven to be a very strong and difficult to improve baseline (Martinez et al., 2017). This is especially true in our use case, where one interlocutor tends to keep a static pose while thinking or listening to the other interlocutor's speech.

5.5. Top-performing solutions

The scores of the challenge participants for each part of the body are reported in Table 6. As can be seen, none of the teams outperformed the zero-velocity baseline. However, the *SAIR KCL* team (Tuyen and Celiktutan, 2021) was awarded with an honorable mention for the novelty of their approach[13], which is presented next.

5.5.1. Honorable mention: SAIR KCL

Motivated by the influence of social signals in communication (Breazeal, 2002), the *SAIR KCL* team proposed a method that exploits the dyadic information to improve the behavior forecasting task. In a sequence-to-sequence fashion (*seq2seq*, Sutskever et al. 2014), three LSTM units encode the 2D landmarks of the target person's observed face, body and hands. Another single LSTM unit followed by a fully connected layer encodes the 2D landmarks from the interacting partner's face, body and hands all together into a fixed-length contextual vector. In the decoding stage, the 2D landmarks from the target person's

13. https://github.com/TuyenNguyenTanViet/ForecastingNonverbalSignals

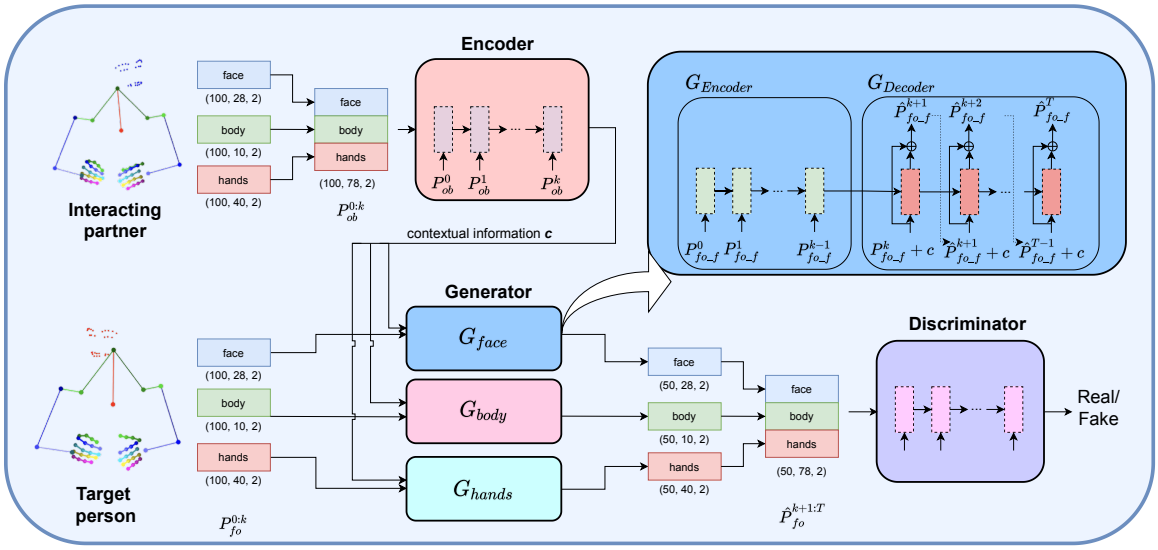

Figure 10: Framework proposed by the *SAIR KCL* team[13] to forecast non-verbal behaviors of the target person during social dyadic interactions. The input consists of the landmarks of the target and the interacting partner observed during 4 seconds (100 frames).

face, body, and hands at frame t are each concatenated to the contextual vector and then fed to their respective LSTM decoder, which generates the predicted pose for the frame *t+1*. Additionally, a residual layer is added between the input and the output of each LSTM cell of the decoder in order to better model the velocity of the motion (Martinez et al., 2017). Finally, they also incorporate a discriminator, which learns to distinguish between the real future sequence of poses and the generated one. They are both trained in an adversarial way so that the generator is encouraged to produce more realistic motions. The approach is depicted in Figure 10.

The authors remark that the low accuracy when predicting hands gestures might be due to their inherent randomness. They suggest that building a multimodal setting to exploit audio might help the network to learn the rhythmic movements (beat gestures) which usually accompany the communicator's speech (McNeill, 2011).

5.5.2. Analysis of results

In an additional experiment, we computed the absolute differences of the predicted landmarks between consecutive frames for all the test segments (Δ). This metric quantifies the amount of motion in the predicted sequences of landmarks. In order to assess the evolution of the predictions motion along the future, this metric was also computed restricted to the short- (<400ms), mid- (400ms to 1s) and long-term (1s to 2s) future windows (Δ_S, Δ_M, and Δ_L, respectively). Finally, we also retrieved the mean absolute differences between the last observed and the first predicted landmarks (Δ_0). Results for the test submissions of both final stage participants are summarized in Table 7. In spite of the fact that the

	Face				Body				Hands			
Team	Δ_0	Δ_S	Δ_M	Δ_L	Δ_0	Δ_S	Δ_M	Δ_L	Δ_0	Δ_S	Δ_M	Δ_L
Baseline	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Ground truth	0.56	0.62	0.66	0.67	0.24	0.29	0.32	0.31	0.59	1.28	1.31	1.36
rays2pix	0.00	0.00	0.00	0.00	0.02	0.45	0.27	0.14	0.90	0.62	0.47	0.40
SAIR KCL	1.99	0.36	0.40	0.53	1.66	0.19	0.15	0.15	4.61	0.48	0.44	0.70

Table 7: Absolute mean differences (in pixels) across test segments between last observed and first predicted landmarks (Δ_0), and between consecutive short- (Δ_S), mid- (Δ_M), and long-term (Δ_L) predictions (0 to 400ms, 400ms to 1s, and 1 to 2s, respectively).

rays2pix team did not submit their fact sheet, we can observe that their method is equivalent to the zero-velocity baseline for face. Their body and hands predictions tend to freeze near the end of the prediction window, especially for the body. This effect is alleviated in *SAIR KCL*'s framework by the adversarial loss, as a progressive deceleration would be a strong discriminative feature for distinguishing real from generated sequences. Not only their predictions do not freeze, but their motion increases when comparing the short- and the long-term windows (from 0.36 to 0.53 and from 0.48 to 0.70 for face and body, respectively). A qualitative analysis would determine whether such acceleration is due to instability in the long-term predictions. Finally, we hypothesize that the high Δ_0 values observed for the honorable mention method in all the body parts could be the origin of its modest challenge performance. High values of Δ_0 translate to an overall rough transition between the observation and the predicted window. Since the discriminator only sees the predicted sequence of landmarks, it does not explicitly enforce a smooth and realistic transition between the observation and the prediction window. As a result, the combination of 1) a model rewarded for generating plausible accurate predictions and 2) the huge stochasticity of the future might favor the generation of realistic pose sequences in an average safe position.

5.6. Discussion

In the Behavior Forecasting track, challenge participants failed to outperform the zero-velocity baseline. Indeed, such baseline has proven to be simple but strong in other motion forecasting scenarios (Martinez et al., 2017). It is even more powerful in our test scenario (i.e., the *Talk* task), where participants often remain static while listening to their partner, or perform very subtle movements (e.g., nodding). Some of these subtle face movements may be better captured with action units and gaze direction information, instead of solely relying on facial landmarks. Added to the high stochasticity of the future, predicting a future sequence of landmarks with high motion is often unworthy in terms of risk versus success rate. As a matter of fact, in the very recent work from Barquero et al. (2022b) on UDIVA v0.5, the zero-velocity baseline has been beaten by methods that predict slowly moving and eventually freezing sequences without adversarial training. As a result, we might

question whether quantitative metrics are an appropriate indicator of good performance by themselves. Similarly, we often find situations where several distinct future behaviors are equally valid and contextually coherent under the same observable circumstances. Such situations are especially frequent when dealing with long-term horizons like ours (2 seconds). In these scenarios, considering only a single observed behavior as the ground truth may be detrimental. A possible solution could be using *human criteria*. Unfortunately, qualitative assessments like questionnaires are unscalable and time-consuming options that are not free of biases either.

Overall, the participants from the Behavior Forecasting track highlighted the difficulty of predicting accurate hands gestures. Although there are no previous works studying hands forecasting in other scenarios to the best of our knowledge, we hypothesize that three main aspects may drive such difficulty. First, the hands gestures are extremely stochastic. In conjunction with the primarily static nature of our scenario, predictive models have little incentive to generate hands gestures over hands remaining in a static and safe position. Second, the commonly used mean squared or derived losses only reward accurate gesture predictions when they are time-synchronized to the ground truth at the frame level. If not, the reward might even become a penalization. Finally, the hands annotations are very noisy and frequently missing due to occlusions, hands interaction, and rapid movement. As a result, the training might become unstable. Behavior forecasting will therefore benefit from the advances in pose estimation methods, especially for hands. In particular, very recent works specialize in interacting hands pose estimation (Zhang et al., 2021a; Kim et al., 2021).

One of the main goals of this challenge was to promote novel architectures that exploited the social dynamics of the interactions. Even though the *SAIR KCL*'s method encodes the observed landmarks from both participants to generate the future motion of the target interlocutor, we lack a comparison against the same method without exploiting the interacting partner's landmarks. Recent works in similar benchmarks showed the potential of considering only one participant of the interaction (Wang et al., 2021a; Barquero et al., 2022b). We hypothesize that the challenges behind successfully exploiting dyadic and context information may be mainly two. On the one hand, a vast amount of data is required to model the complex data distribution of the dyads. On the other hand, the imbalanced ratio of monadic- versus dyad-driven dynamics and interactions with other sources of context in a conversation may impede the network to take full advantage of the minority class. However, the joint computational modeling of both participants is a promising line based on the psychology literature (Kaye and Fogel, 1980; Whyatt and Torres, 2017). In fact, recent works focusing on highly interactive scenarios like dancing have already proved the benefits from exploiting the dyad dynamics (Katircioglu et al., 2021; Guo et al., 2021). Similarly, we also wanted the participants to explore whether the combination of multiple modalities helped to improve the predictions. Unfortunately, none of this track's participants exploited multimodal approaches. Preliminary studies though have recently reported that naively incorporating multimodal data like audio, transcripts, or metadata might not have any significant impact on the results (Barquero et al., 2022b). Consequently, further research is needed in order to find the right multimodal fusion strategies that allow the networks to learn useful multimodal behavioral patterns. We strongly believe that incorporating information from other modalities may help narrow the amount of possible futures

and improve the accuracy of the predictions. Actually, multimodal approaches have already proved useful in other higher-level social signal forecasting problems (Raman et al., 2021; Boudin et al., 2021). In our problem, the main obstacle of incorporating other modalities resides in the natural increase of the problem dimensionality. For example, leveraging the last heard utterances of the participant might help anticipating their next gesture. However, the spectrum of possible combinations of observed landmarks and last utterances is immense, and most surely underrepresented in any available training set. Generally, including all possible combinations of verbal/non-verbal, individual/social behaviors in a dataset is extremely challenging. Therefore, the research community should invest efforts in collecting large-scale unlabeled datasets that capture more varied individual and social behavior representations. We expect such datasets to favor data-hungry strategies like transfer learning, which remains unexplored as of today. This would eventually benefit the deployment of behavior forecasting in scenarios where specific data collection is difficult due to privacy concerns, for example. Finally, we also foresee future work to research on multitask settings where multiple social cues and signals are simultaneously predicted, the benefits of which have already been shown in other behavioral forecasting works (Ishii et al., 2020, 2021).

6. Ethical considerations

The research presented herein has several advantages and applications for good (see Section 4.1 and Section 5.1), but also comes with a number of potential ethical pitfalls. In this section, we describe some of the ethical issues concerning the methodological aspects of the challenge as well as its results and consequences.

Since this research involved the collection and manipulation of sensitive data, several ethical aspects were considered to conduct it. Consent to collect and use the data was asked with full disclosure of how it would be used, processed, and for how long the data would be available for further processing. Additionally, the data is preserved anonymously and encrypted, all participation was voluntary, and it entailed no degree of harm. A noteworthy ethical concern in our research is the sample bias towards a WEIRD population (Henrich, 2020), since most of the participants in our sample matched the characteristics of a white and highly educated population.

On a different note, there are important ethical concerns in the use of the results in personality recognition and behavior forecasting fields. First off, since personality computing may become the new channel for psychological assessment over the next few years (Bleidorn and Hopwood, 2019), researchers and practitioners should ensure data privacy even more as the possibilities of data leakage and misuse have increased (Nayak and Ojha, 2020). Collected personality data could be misused for personality profiling beyond the scopes of the research purposes, like personnel selection (Tippins et al., 2021), selling strategies (Dhelim et al., 2021) or such as it happened in the Cambridge Analytica scandal, when personalized political messages were delivered based on reconstructions of personality profiles from social networks (Confessore, 2018). Similarly, non-consensual behavior forecasting may have potential pitfalls in areas such as security borders or migration controls, where unfair algorithms may lead to undesired outcomes (McKendrick, 2019) impacting human rights (Akhmetova and Harris, 2021). All in all, although data protection regulations vary across countries (Guzzo et al., 2015), we encourage future research to ensure informational

self-determination and consensual use of the information that can be extracted with the methods presented herein. In this sense, frameworks such as the EU GDPR provide excellent safeguards for establishing ethical borders that should not be crossed.

7. Conclusion

This paper summarizes all the research derived from the Personality Recognition and Behavior Forecasting tracks of the ChaLearn LAP DYAD'21 Challenge on Understanding Social Behavior in Dyadic and Small Group Interactions. The paper also describes the UDIVA v0.5 dataset, used for the challenge, and discusses the challenge organization, outcomes, and possible future research lines. Personality recognition and human behavior forecasting are multidisciplinary and complex problems. For the Personality Recognition track, the winning team was able to improve the baseline, showing promising performance and generalization capabilities. Both winning and honorable mention teams showed the importance of taking into account metadata when recognizing self-reported personality data. For the Behavior Forecasting track, even if participants were not able to outperform the zero-velocity baseline, we have raised fundamental questions about the problem. How to deal with noisy and missing data and how to promote realistic but also accurate predictions remain open challenges. Furthermore, the exploitation of multimodal architectures remains unexplored. We expect future research lines to focus on building novel multimodal frameworks that will become the new state of the art for behavior forecasting.

The DYAD challenge (and associated workshop[14]) brought together researchers in the field and from related disciplines (e.g., computer vision, machine learning, affective computing, social signal processing, human-machine interaction, social sciences, social psychology) to discuss the advances and new challenges on the topic of dyadic and small group interactions, explore strengths and limitations of existing approaches, and help define the future directions of the field. We believe this work, in addition to UVIDA and UVIDA v0.5 datasets, may become a reference in the field. The datasets can be further expanded and exploited in many ways, either through additional annotations (e.g., high-level behavior labels, perceived personality state over time) or behavioral analysis from different perspectives.

Acknowledgments

We acknowledge ChaLearn and 4Paradigm for their support on annotating the dataset and for sponsoring the DYAD workshop and challenge. We acknowledge Meta Reality Labs and 4Paradigm for sponsoring the DYAD challenge prizes. We thank all challenge participants and researchers that requested dataset access for their interest in the challenge. This research was supported by Spanish project PID2019-105093GB-I00, ICREA under the ICREA Academia program, and ANR Chair of Artificial Intelligence HUMANIA ANR-19-CHIA-0022.

14. https://chalearnlap.cvc.uab.cat/workshop/44/description/

References

Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters*, 5:6033–6040, 10 2020.

Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Nasir Ahmed, T. Natarajan, and Kamisetty R. Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

Chaitanya Ahuja, Louis Philippe Morency, Yaser Sheikh, and Shugao Ma. To react or not to react: End-to-end visual pose forecasting for personalized avatar during dyadic conversations. *2019 International Conference on Multimodal Interaction*, pages 74–84, 10 2019.

R. Akhmetova and E. Harris. Politics of technology: the use of artificial intelligence by us and canadian immigration agencies and their impacts on human rights. In *Digital Identity, Virtual Borders and Social Media*. Edward Elgar Publishing, 2021.

Ridwan Alam, Azziza Bankole, Martha Anderson, and John Lach. Multiple-instance learning for sparse behavior modeling from wearables: Toward dementia-related agitation prediction. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1330–1333. IEEE, 2019.

Leo Alexander III, Evan Mulfinger, and Frederick L Oswald. Using big data and machine learning in personality measurement: Opportunities and challenges. *European Journal of Personality*, 34(5):632–648, 2020.

Guozhen An and Rivka Levitan. Lexical and Acoustic Deep Learning Model for Personality Recognition. In *Proc. Interspeech 2018*, pages 1761–1765, 2018.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.

Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3603–3612, 2015.

Salvatore Anzalone, Giovanna Varni, Serena Ivaldi, and Mohamed Chetouani. Automated prediction of extraversion during human-humanoid interaction. *Int. Journal of Social Robotics*, 9(3):385–399, 2017.

Oya Aran and Daniel Gatica-Perez. One of a kind: Inferring personality impressions in meetings. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 11–18, 2013.

German Barquero, Johnny Núñez, Sergio Escalera, Zhen Xu, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Didn't see that coming: a survey on non-verbal social human behavior forecasting. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, Proceedings of Machine Learning Research, pages 139–178, 2022a.

German Barquero, Johnny Núñez, Zhen Xu, Sergio Escalera, Wei-Wei Tu, Isabelle Guyon, and Cristina Palmero. Comparison of spatio-temporal models for human motion and pose forecasting in face-to-face interaction scenarios. In *Understanding Social Behavior in Dyadic and Small Group Interactions*, Proceedings of Machine Learning Research, pages 107–138, 2022b.

Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018.

Ligia Batrinca, Bruno Lepri, and Fabio Pianesi. Multimodal recognition of personality during short self-presentations. In *ACM Workshop on Human Gesture and Behavior Understanding*, pages 27–28, 2011.

Wiebke Bleidorn and Christopher James Hopwood. Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2): 190–203, 2019.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. A multimodal model for predicting conversational feedbacks. In *International Conference on Text, Speech, and Dialogue*, pages 537–549. Springer, 2021.

Dawn O Braithwaite and Paul Schrodt. *Engaging theories in interpersonal communication: Multiple perspectives*. Routledge, 2021.

Cynthia L Breazeal. *Designing sociable robots*. MIT press, 2002.

Egon Brunswik. *Perception and the representative design of psychological experiments*. Univ of California Press, 1956.

Isabelle Bülthoff, Wonmo Jung, Regine G. M. Armann, and Christian Wallraven. Predominance of eyes and surface information for face race categorization. *Scientific Reports*, 11 (1):1927, 2021.

Judee K Burgoon, Lesa A Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 1995.

Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, pages 226–242. Springer, 2020.

Géry Casiez, Nicolas Roussel, and Daniel Vogel. 1€ filter: a simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2527–2530, 2012.

Fabio Celli, Elia Bruni, and Bruno Lepri. Automatic personality and interaction style recognition from facebook profile pictures. In *International Conference on Multimedia*, pages 1101–1104, 2014.

Mohamed Chaabane, Ameni Trabelsi, Nathaniel Blanchard, and Ross Beveridge. Looking ahead: Anticipating pedestrians crossing with future frames prediction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020.

Zezhou Chen, Zhaoxiang Liu, Huan Hu, Jinqiang Bai, Shiguo Lian, Fuyuan Shi, and Kai Wang. A realistic face-to-face conversation system based on deep neural networks. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019.

Hsu-kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE, 2019.

Hang Chu, Daiqing Li, and Sanja Fidler. A face-to-face neural conversation model. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*, 2018.

Junyoung Chung, Sungjin Ahn, and Yoshua Bengio. Hierarchical multiscale recurrent neural networks. In *5th International Conference on Learning Representations, ICLR 2017*, 2016.

Nicholas Confessore. Cambridge analytica and facebook: The scandal and the fallout so far. Published at The New York Times 2018, April 4., 2018.

Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5041, 2020a.

Enric Corona, Albert Pumarola, and Guillem Alenyà. Context-aware human motion prediction. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6992–7001, 2020b.

Ronen Cuperman and William Ickes. Big five predictors of behavior and perceptions in initial dyadic interactions: Personality similarity helps extraverts and introverts, but hurts "disagreeables". *Journal of personality and social psychology*, 97(4):667, 2009.

David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B. Moeslund, Sergio Escalera, and Cristina Palmero. Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2177–2188, October 2021.

Kerstin Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480):679–704, 2007.

Robert O. Davis. The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review*, 24:193–219, 2018.

Emilie Delaherche, Mohamed Chetouani, Ammar Mahdhaoui, Catherine Saint-Georges, Sylvie Viaux, and David Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Transactions on Affective Computing*, 3(3):349–365, 2012.

Soumia Dermouche and Catherine Pelachaud. Engagement modeling in dyadic interaction. In *2019 International Conference on Multimodal Interaction*, pages 440–445, 2019.

Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. A survey on personality-aware recommendation systems. *Artificial Intelligence Review*, pages 1–46, 2021.

Peter Ford Dominey, Giorgio Metta, Francesco Nori, and Lorenzo Natale. Anticipation and initiative in human-humanoid interaction. In *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*, pages 693–699. IEEE, 2008.

Nuno Ferreira Duarte, Mirko Raković, Jovica Tasevski, Moreno Ignazio Coco, Aude Billard, and José Santos-Victor. Action anticipation: Reading the intentions of humans and robots. *IEEE Robotics and Automation Letters*, 3(4):4132–4139, 2018.

Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk KJ Heylen, Hayley Hung, Mark A Neerincx, and Khiet P Truong. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 206–212. IEEE, 2019.

Hugo Jair Escalante, Heysem Kaya, Albert Ali Salah, Sergio Escalera, Yagmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio C. S. Jacques Junior, Meysam Madadi, Stéphane Ayache, Evelyne Viegas, Furkan Gürpinar, Achmadnoer Sukma Wicaksana, Cynthia C. S. Liem, Marcel A. J. van Gerven, and Rob van Lier. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, 2020.

Valentín Escudero, Minsun Lee, and Myrna L. Friedlander. *Dyadic Interaction Analysis*, page 45–67. Cambridge Handbooks in Psychology. Cambridge University Press, 2018.

Anna Esposito, Leopoldina Fortunati, and Giuseppe Lugano. Modeling emotion, behavior and context in socially believable robots and ict interfaces. *Cognitive Computation*, 6(4): 623–627, 2014.

Connor Esterwood and Lionel P Robert. A systematic review of human and robot personality in health care human-robot interaction. *Frontiers in Robotics and AI*, page 306, 2021.

European Parliament and Council of European Union. Regulation (eu) 2016/679. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679, 2016.

Sheng Fang, Catherine Achard, and Séverine Dubuisson. Personality classification and behaviour interpretation: An approach based on feature categories. In *ICMI*, pages 225–232, 2016.

Will Feng, Anitha Kannan, Georgia Gkioxari, and C. Lawrence Zitnick. Learn2smile: Learning non-verbal interaction through observation. *IEEE International Conference on Intelligent Robots and Systems*, 2017-September:4131–4138, 12 2017.

William Fleeson and Eranda Jayawickreme. Whole trait theory. *Journal of research in personality*, 56:82–92, 2015.

Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 329–338. Association for Computing Machinery, 2019. ISBN 9781450361255.

Naoya Fushishita, Antonio Tejero-de-Pablos, Yusuke Mukuta, and Tatsuya Harada. Long-term video generation of multiple futures using human poses. *CoRR*, abs/1904.07538, 2019.

David Gallardo-Pujol, Antonio Andrés-Pueyo, and Alberto Maydeu-Olivares. Maoa genotype, social exclusion and aggression: An experimental test of a gene–environment interaction. *Genes, Brain and Behavior*, 12(1):140–145, 2013.

David Gallardo-Pujol, Victor Rouco, Anna Cortijos-Bernabeu, Luis Oceja, Christopher J Soto, and Oliver P John. Factor structure, gender invariance, measurement properties and short forms of the spanish adaptation of the big five inventory-2 (BFI-2). 2021.

Daniel Gatica-Perez. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and vision computing*, 27(12):1775–1787, 2009.

Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Pratik Gujjar and Richard Vaughan. Classifying pedestrian actions in advance using predicted video of urban driving scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2097–2103. IEEE, 2019.

Didem Gundogdu, Ailbhe N Finnerty, Jacopo Staiano, Stefano Teso, Andrea Passerini, Fabio Pianesi, and Bruno Lepri. Investigating the association between social interactions and personality states dynamics. *Royal Society open science*, 4(9):170194, 2017.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 152–168. Springer, 2020.

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Multi-person extreme motion prediction with cross-interaction attention. *arXiv preprint arXiv:2105.08825*, 5 2021.

Richard A Guzzo, Alexis A Fink, Eden King, Scott Tonidandel, and Ronald S Landis. Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology*, 8(4):491–508, 2015.

Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel A.J. van Gerven, and Rob van Lier. Multimodal first impression analysis with deep residual networks. *IEEE Transactions on Affective Computing*, 9(3):316–329, 2018.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3196–3206, 2020.

John A Hartigan. *Clustering algorithms*. John Wiley & Sons, Inc., 1975.

Joseph Henrich. *The WEIRDest people in the world: How the West became psychologically peculiar and particularly prosperous*. Penguin UK, 2020.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. CNN architectures for large-scale audio classification. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 131–135. IEEE, 2017.

Louis Hickman, Rachel Saef, Vincent Ng, Sang Eun Woo, Louis Tay, and Nigel Bosch. Developing and evaluating language-based machine learning algorithms for inferring applicant personality in video interviews. *Human Resource Management Journal*, 2021.

Yutaro Honda, Rei Kawakami, and Takeshi Naemura. Rnn-based motion prediction in competitive fencing considering interaction between players. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press, 2020.

Sabrina Hoppe, Tobias Loetscher, Stephanie A Morey, and Andreas Bulling. Eye movements during everyday behavior predict personality traits. *Frontiers in human neuroscience*, page 105, 2018.

Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *CoRR*, abs/1704.04861, 2017.

Minjie Hua, Fuyuan Shi, Yibing Nan, Kai Wang, Hao Chen, and Shiguo Lian. Towards more realistic human-robot conversation: A seq2seq-based body gesture interaction system. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1393–1400, 2019.

Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pages 83–90. IEEE, 2016.

Yuchi Huang and Saad M Khan. Dyadgan: Generating facial expressions in dyadic interactions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.*, 2017.

Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. Can prediction of turn-management willingness improve turn-changing modeling? In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–8, 2020.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 131–138, 2021.

Julio C. S. Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel AJ Van Gerven, Rob Van Lier, et al. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, 2019.

Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Baro, and Sergio Escalera. Person perception biases exposed: Revisiting the first impressions dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 13–21, January 2021.

Haifeng Jin, Qingquan Song, and Xia Hu. Auto-keras: An efficient neural architecture search system. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 1946–1956, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362016.

Kacper Kania, Marek Kowalski, and Tomasz Trzciński. Trajevae–controllable human motion generation from trajectories. *arXiv preprint arXiv:2104.00351*, 2021.

Isinsu Katircioglu, Costa Georgantas, Mathieu Salzmann, and Pascal Fua. Dyadic human motion prediction. *arXiv preprint arXiv:2112.00396*, 12 2021.

Kenneth Kaye and Alan Fogel. The temporal structure of face-to-face communication between mothers and infants. *Developmental psychology*, 16(5):454, 1980.

Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11189–11198, 2021.

Ue-Hwan Kim, Dongho Ka, Hwasoo Yeo, and Jong-Hwan Kim. A real-time vision framework for pedestrian behavior recognition and intention prediction at intersections using 3d pose estimation. *arXiv preprint arXiv:2009.10868*, 2020.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Jangwon Lee, Haodan Tan, David Crandall, and Selma Šabanović. Forecasting hand gestures for human-drone interaction. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 167–168, 2018.

Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. Employing social gaze and speaking activity for automatic determination of the extraversion trait. In *International conference on multimodal interfaces and the workshop on machine learning for multimodal interaction*, pages 1–8, 2010.

Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. Connecting meeting behavior with extraversion—a systematic study. *IEEE Transactions on Affective Computing*, 3(4):443–455, 2012a.

Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. Connecting meeting behavior with extraversion—a systematic study. *IEEE Transactions on Affective Computing*, 3(4):443–455, 2012b.

Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019.

Yun-Shao Lin and Chi-Chun Lee. Using interlocutor-modulated attention blstm to predict personality traits in small group interaction. In *International Conference on Multimodal Interaction*, pages 163–169, 2018.

Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020.

Herb Marsh, Benjamin Nagengast, and Alexandre Morin. Measurement invariance of big-five factors over the life span: Esem tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental psychology*, 49, 01 2012.

Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Robert R. McCrae and Oliver P. John. An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2):175–215, 1992.

Kathleen McKendrick. Artificial intelligence prediction and counterterrorism. *London: The Royal Institute of International Affairs-Chatham House*, 9, 2019.

David McNeill. *Hand and mind*. De Gruyter Mouton, 2011.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Yash Mehta, Navonil Majumder, Alexander Gelbukh, and Erik Cambria. Recent trends in deep learning based personality detection. *Artificial Intelligence Review*, pages 1–27, 2019.

Yaw M Mensah and Hsiao-Yin Chen. Global clustering of countries by culture–an extension of the globe study. *Available at SSRN 2189904*, 2013.

Philip Moore. Do we understand the relationship between affective computing, emotion and context-awareness? *Machines*, 5(3):16, 2017.

Amir Nadeem, Ahmad Jalal, and Kibum Kim. Human actions tracking and recognition based on body parts detection via artificial neural network. In *2020 3rd International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–6. IEEE, 2020.

Suvendu Kumar Nayak and Ananta Charan Ojha. Data leakage detection and prevention: Review and research directions. *Machine Learning and Information Processing*, pages 203–212, 2020.

Laurent Son Nguyen, Alvaro Marcos-Ramiro, Martha Marrón Romera, and Daniel Gatica-Perez. Multimodal analysis of body communication cues in employment interviews. In *International Conference on Multimodal Interaction (ICMI)*, pages 437–444, 2013.

Olivia Nocentini, Laura Fiorini, Giorgia Acerbi, Alessandra Sorrentino, Gianmaria Mancioppi, and Filippo Cavallo. A survey of behavioral models for social robots. *Robotics*, 8 (3):54, 2019.

Shogo Okada, Oya Aran, and Daniel Gatica-Perez. Personality trait classification via co-occurrent multiparty multimodal event discovery. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 15–22, 2015.

Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3): 1–40, 2017.

Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapés, Alexa Moseguí, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, and Sergio Escalera. Context-aware personality inference in dyadic scenarios: Introducing the udiva dataset. In *2021 IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 1–12, 2021.

Behnam Parsaeifard, Saeed Saadatnejad, Yuejiang Liu, Taylor Mordan, and Alexandre Alahi. Learning decoupled representations for human pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2303, 2021.

Le Vy Phan and John F Rauthmann. Personality computing: New frontiers in personality assessment. *Social and Personality Psychology Compass*, 15(7):e12624, 2021.

Fabio Pianesi, Nadia Mana, Alessandro Cappelletti, Bruno Lepri, and Massimo Zancanaro. Multimodal recognition of personality traits in social interactions. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 53–60, 2008.

Rosalind W Picard. *Affective computing*. MIT press, 2000.

Ricardo Darío Pérez Principi, Cristina Palmero, Julio C. S. Jacques Junior, and Sergio Escalera. On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, 12(3):607–621, 2019.

Jianing Qiu, Frank P-W Lo, Xiao Gu, Yingnan Sun, Shuo Jiang, and Benny Lo. Indoor future person localization from an egocentric wearable camera. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8586–8592. IEEE, 2021.

Chirag Raman, Hayley Hung, and Marco Loog. Social processes: Self-supervised forecasting of nonverbal cues in social conversations. *arXiv preprint arXiv:2107.13576*, 2021.

John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. The situational eight diamonds: A taxonomy of major dimensions of situation characteristics. *Journal of Personality and Social Psychology*, 107(4):677, 2014.

Brent W Roberts, Kate E Walton, and Wolfgang Viechtbauer. Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological bulletin*, 132(1):1, 2006.

Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021.

Giovanni Saponaro, Giampiero Salvi, and Alexandre Bernardino. Robot anticipation of human intentions through continuous gesture recognition. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 218–225. IEEE, 2013.

István Sárándi, Timm Linder, Kai O Arras, and Bastian Leibe. Metrabs: Metric-scale truncation-robust heatmaps for absolute 3d human pose estimation. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

Felix D Schönbrodt and Marco Perugini. At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5):609–612, 2013.

Kevin Sexton, Amanda Johnson, Amanda Gotsch, Ahmed A Hussein, Lora Cavuoto, and Khurshid A Guru. Anticipation, teamwork and cognitive load: chasing efficiency during robot-assisted surgery. *BMJ quality & safety*, 27(2):148–154, 2018.

Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020.

Zilong Shao, Siyang Song, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. Personality recognition by modelling person-specific cognitive processes using graph representation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 357–366, 2021.

Tianmin Shu, M. S. Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. *IJCAI International Joint Conference on Artificial Intelligence*, 2016-January:3454–3461, 4 2016.

Theodoros Sofianos, Alessio Sampieri, Luca Franco, and Fabio Galasso. Space-time-separable graph convolutional network for pose forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11209–11218, 2021.

Gizem Sogancioglu, Heysem Kaya, and Albert Ali Salah. Can mood primitives predict apparent personality? In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

Siyang Song, Zilong Shao, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. Learning graph representation of person-specific cognitive processes from audio-visual behaviours for automatic personality recognition. *arXiv preprint arXiv:2110.13570*, 2021.

Christopher Soto and Oliver John. The next big five inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology*, 113:117–143, 07 2017.

Clemens Stachl, Quay Au, Ramona Schoedel, Samuel D Gosling, Gabriella M Harari, Daniel Buschek, Sarah Theres Völkel, Tobias Schuwerk, Michelle Oldemeier, Theresa Ullmann, et al. Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30):17680–17687, 2020.

Alexandros Stergiou and Ronald Poppe. Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding*, 188:102799, 2019.

Ming-Hsiang Su, Chung-Hsien Wu, and Yu-Ting Zheng. Exploiting turn-taking temporal evolution for personality trait perception in dyadic conversations. *Transactions on Audio, Speech, and Language Processing*, 24(4):733–744, 2016.

Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Oswald Lanz, and Nicu Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *ICMI*, pages 3–10, 2013.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Yongyi Tang, Lin Ma, Wei Liu, and Wei-Shi Zheng. Long-term human motion prediction by modeling motion context and enhancing motion dynamics. In *IJCAI*, pages 935–941, 2018.

Adriana Tapus and Maja J Matarić. User personality matching with a hands-off robot for post-stroke rehabilitation therapy. In *Experimental robotics*, pages 165–175. Springer, 2008.

Nancy T Tippins, Frederick L Oswald, and S Morton McPhail. Scientific, legal, and ethical concerns about ai-based personnel selection tools: a call to action. *Personnel Assessment and Decisions*, 7(2):1, 2021.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6450–6459, 2018.

Nguyen Tan Viet Tuyen and Oya Celiktutan. Forecasting nonverbal social signals during dyadic interactions with generative adversarial neural networks. *arXiv preprint arXiv:2110.09378*, 2021.

Ryosuke Ueno, Yukiko I. Nakano, Jie Zeng, and Fumio Nihei. Estimating the intensity of facial expressions accompanying feedback responses in multiparty video-mediated communication. *ICMI 2020 - Proceedings of the 2020 International Conference on Multimodal Interaction*, 20:144–152, 10 2020.

Fabio Valente, Samuel Kim, and Petr Motlicek. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus. In *Thirteenth annual conference of the international speech communication association*, 2012.

Michel Valstar, Brais Martinez, Xavier Binefa, and Maja Pantic. Facial point detection using boosted regression and graph models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2729–2736. IEEE, 2010.

Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transaction on Affective Computing*, 5(3):273–291, 2014a.

Alessandro Vinciarelli and Gelareh Mohammadi. More personality in personality computing. *IEEE Transactions on Affective Computing*, 5(3):297–300, 2014b.

Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2011.

Alessandro Vinciarelli, Anna Esposito, Elisabeth André, Francesca Bonin, Mohamed Chetouani, Jeffrey F Cohn, Marco Cristani, Ferdinand Fuhrmann, Elmer Gilmartin, Zakia Hammal, et al. Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7 (4):397–413, 2015.

Kathan Vyas, Rui Ma, Behnaz Rezaei, Shuangjun Liu, Michael Neubauer, Thomas Ploetz, Ronald Oberleitner, and Sarah Ostadabbas. Recognition of atypical behavior in autism diagnosis from video using pose estimation over time. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2019.

Chenxi Wang, Yunfeng Wang, Zixuan Huang, and Zhiwen Chen. Simple baseline for single human motion forecasting. *ICCV SoMoF Workshop*, 2021a.

Jiashun Wang, Huazhe Xu, Medhini Narasimhan, Xiaolong Wang, and Uc San Diego. Multi-person 3d motion prediction with multi-range transformers. *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021b.

Yanna Weisberg, Colin Deyoung, and Jacob Hirsh. Gender differences in personality across the ten aspects of the big five. *Frontiers in psychology*, 2:178, 08 2011.

Caroline P Whyatt and Elizabeth B Torres. The social-dance: decomposing naturalistic dyadic interaction dynamics to the'micro-level'. In *Proceedings of the 4th international conference on movement computing*, pages 1–8, 2017.

Kipling D Williams, Cassandra L Govan, Vanessa Croker, Daniel Tynan, Maggie Cruickshank, and Albert Lam. Investigations into differences between social-and cyberostracism. *Group dynamics: Theory, research, and practice*, 6(1):65, 2002.

Jieyeon Woo, Catherine Pelachaud, and Catherine Achard. Creating an interactive human/agent loop using multimodal recurrent neural networks. *WACAI*, 2021.

Aidan GC Wright. Current directions in personality science and the potential for advances through computing. *IEEE Transactions on Affective Computing*, 5(3):292–296, 2014.

Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2018.

Ye Yuan and Kris Kitani. Dlow: Diversifying latent flows for diverse human motion prediction. In *European Conference on Computer Vision*, pages 346–364. Springer, 2020.

Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11354–11363, 2021a.

Lingyu Zhang, Indrani Bhattacharya, Mallory Morgan, Michael Foley, Christoph Riedl, Brooke Welles, and Richard Radke. Multiparty visual co-occurrences for estimating personality traits in group meetings. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2085–2094, 2020a.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A cpu real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017.

Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020b.

Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021b.

Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*, pages 218–233. Springer, 2016.