Context-Aware Personality Inference in Dyadic Scenarios: Introducing the UDIVA Dataset

Cristina Palmero[1,2*], Javier Selva[1,2*], Sorina Smeureanu[1,2*], Julio C. S. Jacques Junior[2,3], Albert Clapés[1,2], Alexa Moseguí[1], Zejian Zhang[1,2], David Gallardo[1], Georgina Guilera[1], David Leiva[1], Sergio Escalera[1,2]
[1]Universitat de Barcelona [2]Computer Vision Center [3]Universitat Oberta de Catalunya
{crpalmec7, ssmeursm28, zzhangzh45}@alumnes.ub.edu, jaselvaca@ub.edu,
jsilveira@uoc.edu, aclapes@cvc.uab.cat, alexa.moseguis@gmail.com,
{david.gallardo, gguilera, dleivaur}@ub.edu, sergio@maia.ub.es

Abstract

This paper introduces UDIVA, a new non-acted dataset of face-to-face dyadic interactions, where interlocutors perform competitive and collaborative tasks with different behavior elicitation and cognitive workload. The dataset consists of 90.5 hours of dyadic interactions among 147 participants distributed in 188 sessions, recorded using multiple audiovisual and physiological sensors. Currently, it includes sociodemographic, self- and peer-reported personality, internal state, and relationship profiling from participants. As an initial analysis on UDIVA, we propose a transformer-based method for self-reported personality inference in dyadic scenarios, which uses audiovisual data and different sources of context from both interlocutors to regress a target person's personality traits. Preliminary results from an incremental study show consistent improvements when using all available context information.

1. Introduction

Human interaction has been a central topic in psychology and social sciences, aiming at explaining the complex underlying mechanisms of communication with respect to cognitive, affective, and behavioral perspectives [13, 12]. From a computational point of view, research in dyadic and small group interactions enables the development of automatic approaches for detection, understanding, modeling, and synthesis of individual and interpersonal social signals and dynamics [79]. Many human-centered applications for good (e.g., early diagnosis and intervention [27], augmented telepresence [3], and personalized agents [29]) strongly depend on devising solutions for such tasks.

In dyadic interactions, we use verbal and nonverbal communication channels to convey our goals and intentions [58, 78] while building a common ground [19]. Both interlocutors influence each other based on the cues we perceive [13]. However, the way we perceive, interpret, react, and adapt to them depends on a myriad of factors. Such factors, which we refer to as context, may include, but are not limited to: our personal characteristics, either stable (e.g., personality [21], cultural background, and other sociodemographic information [69]) or transient (e.g., mood [20], physiological or biological factors); the relationship and shared history between both interlocutors; the characteristics of the situation and task at hand; societal norms; and environmental factors (e.g., temperature). What is more, to analyze individual behaviors during a conversation, the joint modeling of both interlocutors is required due to the existing dyadic interdependencies. While these aspects are usually contemplated in non-computational dyadic research [41], context- and interlocutor-aware computational approaches are still scarce, largely due to the lack of datasets providing contextual metadata in different situations and populations [26].

Here, we introduce UDIVA, a highly varied multimodal, multiview dataset of zero- and previous-acquaintance, face-to-face dyadic interactions. It consists of 188 interaction sessions, where 147 participants arranged in dyads performed a set of tasks in different circumstances in a lab setting. It has been collected using multiple audiovisual and physiological sensors, and currently includes sociodemographic, self- and peer-reported personality, internal state, and relationship profiling. To the best of our knowledge, there is no similar publicly available, face-to-face dyadic dataset in the research field in terms of number of views, participants, tasks, recorded sessions, and context labels.

As an initial analysis on the UDIVA dataset, we also propose a novel method for self-reported personality inference in dyadic scenarios. Apart from its importance in interaction understanding, personality recognition is key to develop individualized, empathic, intelligent sys-

*These authors contributed equally to this work.

tems [61]. Our proposal is based on the Video Action Transformer [34], which classifies people's actions in a video by taking advantage of the spatiotemporal context around them. Inspired by [66], we extend query, key, and value from [34] with the other interlocutor's scene, audio, and further context metadata. The latter includes stable and transient characteristics from each interlocutor, as well as session, task, and relationship information. Finally, we experimentally evaluate the usefulness of each additional input incrementally, showing consistent improvements when using all the available context sources and modalities.

2. Related work

This section reviews related work on dyadic scenarios along three axes: social signals and behaviors in context, personality recognition, and human interaction datasets.

Social signals and behaviors in context. Dyadic interactions are a rich source of overt behavioral cues. They can reveal our personal attributes and cognitive/affective inner states dependent upon the context in which they are situated. Context can take many forms, and in the case of recognition of an individual state or attribute, the interaction partner's attributes and behaviors can be considered part of the target person's context. From a computational perspective, spatiotemporal and multiview information can be referred to as context as well. For the measurement of interpersonal constructs (e.g., synchrony [22], rapport [87]), individual social behaviors (e.g., engagement [23]) and impressions (e.g., dominance [86], empathy [61]), the joint modeling of both interlocutors and/or other sources of context has been frequently considered. However, for the task of recognizing individual attributes such as emotion and personality, context has often been misrepresented, despite recurrent claims on its importance [9, 81, 76, 56].

Recent years have seen a small surge in interlocutor-aware approaches for utterance- or turn-based emotion recognition in conversation [64] and sentiment analysis. Early works were based on handcrafted nonverbal, spatiotemporal dyadic features [44, 54]. Nowadays, most approaches rely on deep learning, using conversation transcripts as input with contextualized word or speaker embeddings [47] and considering past and/or future parts of the conversation as additional context. Temporal modeling of those feature representations has been widely performed via recurrent approaches [50], and more recently with BERT/Transformer-like architectures [88, 46]. Some works have further proposed to enrich models with additional modalities, such as raw audiovisual data to enhance the representation of interlocutors' influences and dynamics [83, 36], or speech cues in addition to the personality of the target speaker [45]. Context-aware personality recognition has followed a similar trend as for emotion, but the literature is even scarcer. We discuss it next.

Automatic personality recognition. Personality is widely defined as the manifestation of individual differences in patterns of thought, feeling, and behavior, that remain relatively stable during time [70]. In the personality computing field [77], it is usually characterized by the basic Big Five traits [51] (*Openness to Experience*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*), often referred to as OCEAN, based on self-reported assessments. Most works focus on personality recognition from the individual point of view, even in a dyadic or small group conversational context [6], using only features from the target person. Preliminary studies tended to use handcrafted features from gestures and speech [59], while more recent works rely on deep learning approaches from raw data [52].

To our knowledge, few methods propose interlocutor- or context-aware methods for personality recognition. The work of [72] was one of the first, leveraging turn-taking temporal evolution from transcript features but focusing on apparent personality recognition (i.e., personality reported by external observers [39]). With respect to self-reported personality inference in small group interactions, [30] regressed individual and dyadic features of personality and social impressions utilizing handcrafted descriptors of prosody, speech and visual activity. Later, [48] proposed an interlocutor-modulated recurrent attention model with turn-based acoustic features. Finally, [85] predicted personality and performance labels by correlation analysis of co-occurrent key action events, which were extracted from head and hand pose, gaze and motion intensity features. Regarding context, just one previous approach added person metadata (e.g., gender, age, ethnicity, and perceived attractiveness) to audiovisual data [65]. However, their goal was to better approximate the crowd biases for apparent personality recognition in one-person videos. Contrary to previous works, we use different sources of context, including both interlocutors, scene, and task information to infer personality, using for the first time a video-based transformer adapted to include audio and further context as metadata.

Human interaction datasets. Research on human behavior and communication understanding has fostered the creation of a plethora of human interaction datasets [42, 62, 26, 71]. Here, we focus on publicly available datasets containing at least audiovisual data, which enable the fusion of multiple modalities and the creation of more complete representations. In the literature, we can find examples of rich, non-acted datasets focused on computer-mediated dyadic settings [16, 43], face-to-face triadic [40, 17], or small group interactions [5]. A number of TV-based datasets with acted interactions also exist [63]. However, in such cases, the interlocutors' internal states are artificially built.

One of the advantages of face-to-face settings is that the full overt behavioral spectrum can be observed and modeled. Existing publicly available face-to-face dyadic inter-

Table 1. Publicly available audiovisual human-human (face-to-face) dyadic interaction datasets. "Interaction", *Acted* (actors improvising and/or following an interaction protocol, i.e. given topics/stimulus/tasks), *Acted** (Scripted), *Non-acted* (natural interactions in lab environment) or *Non-acted** (non-acted but guided by interaction protocol); "F/M", number of participants per gender (Female/Male) or number of participants if gender is not informed; "Sess", number of sessions; "Size", hours of recordings;"#Views", number of RGB cameras used, and *D* is RGB+D, *E* is Ego, *M* is Monochrome. The ϕ symbol is used to indicate missing/incomplete/unclear information on the source.

Name / Year	Focus	Interaction	Modality	Annotations	F/M	Sess	Size	#Views	Lang.
IEMOCAP [14], 2008	Emotion recognition	Acted* & Acted	Audiovisual, face & hands MoCap.	Emotions, transcripts, turn-taking	5/5	5	$\sim$12h	2	English
CID [11], 2008	Speech & conversation analysis	Non-acted & Non-acted*	Audiovisual	Speech features, transcripts	10/6	8	8h	1	French
HUMAINE† [24, 25], 2011	Emotion analysis	Non-acted*	Audiovisual	Emotions	34	18	$\sim$12h	4	English
MMDB [67], 2013	Adult-infant interaction analysis	Non-acted*	Audiovisual, depth, physiological	Social cues (gaze, vocal affects, gestures...)	121	160	$\sim$13.3h	8 + 1D	English
MAHNOB [10], 2015	Mimicry	Non-acted*	Audiovisual, head MoCap.	Head, face and hand gestures, personality scores (self-reported)	29/31	54	11.6h	2 + 13M	English
MIT Interview [57], 2015	Hirability analysis	Non-acted*	Audiovisual	Hirability, speech features, social & behavioral traits, transcripts	43/26	138	10.5h	2	English
Creative IT [55], 2016	Emotion recognition	Acted	Audiovisual, body MoCap.	Transcripts, speech features, emotion	9/7	8	$\sim$1h	2	English
MSP-IMPROV [15], 2017	Emotion recognition	Acted & Non-acted	Audiovisual	Turn-taking, emotion	6/6	6	9h	2	English
DAMI-P2C [18], 2020	Adult-infant interaction analysis	Non-acted*	Audiovisual	Emotion, sociodemographics, parenting assessment, child personality (peer-reported)	38/30	65	$\sim$21.6h	1 $^{\phi}$	English
UDIVA (ours), 2020	Social interaction analysis	$\frac{1}{5}$ Non-acted & $\frac{4}{5}$ Non-acted*	Audiovisual, heart rate	Personality scores (self- & peer-reported), sociodemographics, mood, fatigue, relationship type	66/81	188×5 (tasks)	90.5h	6 + 2E	Spanish, Catalan, English

† Here we consider the Green Persuasive and the EmoTABOO [84] databases together.

action datasets are summarized in Table 1[1]. As it can be seen, most of them are limited in the number of participants, recordings, views, context annotations, language, or purpose. The UDIVA dataset has been designed with a multipurpose objective and aims at filling this gap.

3. UDIVA dataset

This section introduces the UDIVA dataset (Understanding Dyadic Interactions from Video and Audio signals), consisting of time-synchronized multimodal, multiview videos of non-scripted face-to-face dyadic interactions based on free and structured tasks performed in a lab setup[2].

3.1. Motivation

UDIVA wants to move beyond automatic individual behavior detection and focus on the development of automatic approaches to study and understand the mechanisms of influence, perception and adaptation to verbal and nonverbal social signals in dyadic interactions, taking into account individual and dyad characteristics as well as other contextual factors. One of our research questions centers on the feasibility of developing systems able to unravel the personality and internal processes of an individual by the social signals they convey, and to understand how interaction partners perceive and react to those cues directed to them. By publicly releasing the dataset to the research community, we encourage data sharing and collaboration among different disciplines, reuse, and repurposing of new research questions.

3.2. Main statistics

The dataset is composed of 90.5h of recordings of dyadic interactions between 147 voluntary participants[3] (55.1% male) from 4 to 84 years old (mean=31.29), coming from 22 countries (68% from Spain). The majority of participants were students (38.8%), and identified themselves as white (84.4%). Participants were distributed into 188 dyadic sessions, with a participation average of 2.5 sessions/participant (max. 5 sessions). To create the dyads, three variables were taken into account: 1) gender (*Female*, *Male*); 2) age group (*Child*: 4-18, *Young*: 19-35, *Adult*: 36-50, and *Senior*: 51-84); and 3) relationship among interlocutors (*Known*, *Unknown*). Participants were matched according to their availability and language while trying to enforce a close-to-uniform distribution among all possible combinations between variables (60 combinations). A minimum age of 4 years and the ability to maintain a conversation in English, Spanish or Catalan were the only inclusion criteria. In the end, the most common interaction group is *Male-Male*/*Young-Young*/*Unknown* (15%), with 43% of the interactions happening among known people. Spanish is the majority language of interaction (71.8%), followed by Catalan (19.7%). Half of the sessions include both interlocutors with Spain as country of origin.

3.3. Questionnaire-based assessments

Prior to their first session, each participant filled a sociodemographic questionnaire, including: age, gender, eth-

[1]The complete table is included in the supplementary material.

[2]Additional details regarding design, participants recruitment, technical setup, and descriptive statistics will be provided in a follow-up paper.

[3]Participants gave consent to be recorded and to share their collected data for research purposes, in compliance with GDPR `https://ec.europa.eu/info/law/law-topic/data-protection_en`.

(a) (b)

Figure 1. Recording environment. We used six tripod-mounted cameras, namely **GB**: General Rear camera, **GF**: General Frontal camera, **HA**: individual High Angle cameras and **FC**: individual Frontal Cameras, and two ego cameras **E** (one per participant, placed around their neck). a) Position of cameras, general microphone and participants. b) Example of the time-synchronized 8 views.

Figure 2. Examples of the 5 tasks included in the UDIVA dataset from 5 sessions. From left to right: *Talk*, *Lego*, *Animals*, *Ghost*, *Gaze*.

nicity, occupation, maximum level of education, and country of origin. To assess personality and/or temperament, age-dependent standardized questionnaires were administered. In particular, parents of children up to 8 years old completed the Children Behavior Questionnaire (CBQ) [68, 60], participants from 9 to 15 years old completed the Early Adolescent Temperament Questionnaire (EATQ-R) [28], while participants aged 16 and older completed both the Big Five Inventory (BFI-2) [70] and the Honesty-Humility axis of the HEXACO personality inventory [8].

All participants (or their parents) completed pre- and post-session mood ([32]) and fatigue (ad hoc 1-to-10 rating scale) assessments. The mood assessment contained items drawn from the Post Experimental Questionnaire of Primary Needs (PEQPN [80]). After each session, participants aged 9 and above completed again the previous temperament/personality and mood questionnaires, this time rating the individual they interacted with, to provide their perceived impression. Finally, participants reported the relationship they had with their interaction partner, if any.

3.4. Structure of a dyadic session

Participants were asked to sit at 90° to one another around a table (see Fig. 1(a)), to be close enough to perform the administered tasks while facilitating data acquisition. A session consisted of 5 tasks (illustrated in Fig. 2) eliciting distinct behaviors and cognitive workload:

Talk. Participants were instructed to talk about any subject during approx. 5 minutes. This task allows analysis of common conversation constructs, such as turn-taking, synchrony, empathy and quality of interaction, among others.

"Animals" game. Participants asked 10 *yes*/*no* questions each to guess the animal they had on their forehead. Animals were classified into 3 difficulty levels. This game reveals cognitive processes (e.g., thinking, gaze events).

Lego building. Participants built a Lego together following the instructions leaflet, ranging between 4 difficulty levels. This task fosters collaboration, cooperation, joint attention, and leader-follower behaviors, among others.

"Ghost blitz" card game. Participants had to select, from a set of 5 figures, the one whose color and shape was not shown in a selected card. They played 1 card per turn, competing with each other to be the first at selecting the correct figure. This task fosters competitive behavior, and allows cognitive processing speed analysis, among others.

Gaze events. Participants followed directions to look *at other's face*, *at static/moving object*, or *elsewhere*, while moving head and eyes. This task serves as ground truth for gaze gestures and face modeling with varied head poses.

These tasks were selected along with psychologists due to the variety of individual and dyadic behaviors they elicit. In particular, Lego structures have been widely used in observational settings to assess aspects as communication [1], social skills [49] or teamwork abilities and performance [31]. *Ghost* and *Animals* are examples of board games, proven to be valid assessments of interpersonal skills [35, 74]. All these aspects are, in turn, indicators of personality traits like *Extraversion*, *Agreeableness* or *Conscientiousness* [4]. Cognitive methods, such as the tasks herein used, are routinely used in personality research [7].

The tasks were explained by a lab proctor prior to each task, who left the recording room while it was taking place. Only for *Gaze* the proctor gave the instructions while participants performed them. *Talk* was always administered first as a warm-up, while *Gaze* was always last. The rest were administered randomly. The difficulty of *Lego* and *Animals*

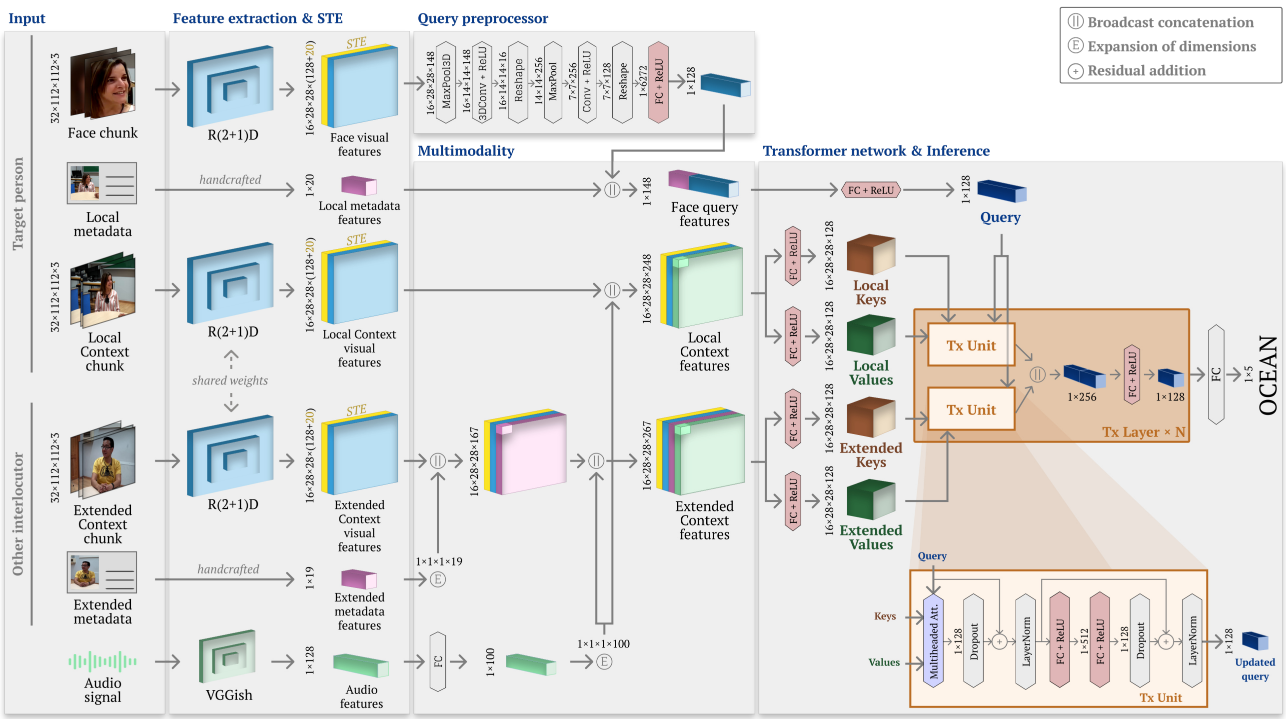

Figure 3. Proposed method to infer self-reported personality (OCEAN) traits from multimodal synchronized signals and context. Input consists of visual (face, local context, and extended context chunks), audio (raw chunks), and metadata (both interlocutors' characteristics, and session and dyadic features). Feature extraction is performed by a R(2+1)D network for the visual chunks and VGGish for audio. The visual features from the R(2+1)D's 3rd residual block are concatenated to spatiotemporal encodings (STE). The VGGish's audio features and handcrafted metadata features are incorporated to visual context/query features and the result transformed to the set of Query, Keys, and Values as input to the Transformer network (Tx). The latter consists of N Tx layers, each equipped with Local and Extended Context Transformer Tx units. Such units implement multiheaded attention and provide their updated queries, which are combined and fed to the next Tx Layer. Finally, the output of the N-th Tx layer is fed to a fully-connected (FC) layer to regress per-chunk OCEAN scores.

for each session was selected such that no participants repeated the same Lego or animal twice, while forcing a uniform distribution on the number of times each item was used for the total of sessions. To assess their difficulty level, we conducted an anonymous survey among 19 co-researchers.

3.5. Technical setup

The setup consisted of 6 HD tripod-mounted cameras (1280×720px, 25fps), 1 lapel microphone per participant and an omnidirectional microphone on the table, as depicted in Fig. 1(a). Each participant also wore an egocentric camera (1920×1080px, 30fps) around their neck and a heart rate monitor on their wrist. All the capturing devices are time-synchronized and the tripod-mounted cameras calibrated. See Fig. 1(b) for an example of the camera views.

4. Personality traits inference

This section provides a first insight into the UDIVA dataset by evaluating it in a personality traits inference task. We present a transformer-based context-aware model to regress self-reported personality traits of a target person during a dyadic interaction. Then, we assess its performance and the effect of adding several sources of context. Method, evaluation protocol and results are described next[4].

4.1. Method description

The attention mechanism of our transformer-based method relates an initial query, in this case the target person's face, to the nonverbal behavior of both interlocutors, the overall scene, and further contextual metadata, and updates it with relevant context. The process is repeated with the updated query in consecutive layers to eventually infer the personality (OCEAN) traits. The proposed method consists of several stages, detailed below. All components and the information flow among them are illustrated in Fig. 3.

Audiovisual input. Let $\mathbf{X}_L, \mathbf{X}_E \in [0, 255]^{\mathcal{L} \times H \times W \times 3}$ be the pair of time-synchronized full-length videos containing the target person (*local context*) and the other interlocutor (*extended context*), respectively. We divide them into 32-frame non-overlapping chunks and resize each chunk's spatial resolution to 112×112 to obtain, respectively, $\mathbf{B}_L, \mathbf{B}_E \in [0, 255]^{32 \times 112 \times 112 \times 3}$. The 32 frames of the chunks are sampled with a stride of 2, such that a

[4] Additional details are provided in the supplementary material.

chunk corresponds to approx. 2.5 seconds of the original videos. Also, we detect the target person's face regions in $\mathbf{X}_L$, crop, and re-scale them to form the face chunk $\mathbf{B}_F \in [0, 255]^{32\times112\times112\times3}$. As face detector, we use a MobileNet-SSD [38] model pretrained on Widerface [82]. Apart from the visual data, we define an audio chunk $\mathbf{b}_A \in \mathbb{R}^{132\,300}$ consisting of the raw audio frames acquired at 44.1 KHz from the general microphone (or one of the lapels if the general one was not available for that session), and time-synchronized to its respective video chunk.

Metadata input. Different sources of context are captured in the form of input metadata, described in Table 2. We consider 2 types of metadata: (1) *local metadata*, containing *individual* context from the target person and *session* information; and (2) *extended metadata*, with *individual* context from the other interlocutor and *dyadic* features.

Feature extraction. First, we normalize the pixel values of $\{\mathbf{B}_F, \mathbf{B}_L, \mathbf{B}_E\}$ in the range $[0, 1]$, subtracting and dividing them by the mean and standard deviation of the IG-65M dataset [33]. Then, we feed them to a R(2+1)D network [73] backbone, pretrained on that same dataset, and save the rich spatiotemporal features produced by the R(2+1)D's 3rd convolutional residual stack: $\mathbf{Z}'_F = g_F(\mathbf{B}_F; \theta_F)$, $\mathbf{Z}'_L = g_L(\mathbf{B}_L; \theta_C)$, $\mathbf{Z}'_E = g_E(\mathbf{B}_E; \theta_C)$, where θ_F are the weights of the face network $g_F(\cdot)$, and θ_C are shared weights of $g_L(\cdot)$ and $g_E(\cdot)$ networks. $\mathbf{Z}'_F, \mathbf{Z}'_L, \mathbf{Z}'_E \in \mathbb{R}^{16\times28\times28\times128}$ denote the *face*, *local context*, and *extended context visual features*, respectively. For the audio feature extraction, we use the VGGish [37] backbone. This VGG-like model, developed specifically for the audio modality and with pretrained weights θ_A learned on a preliminary version of the YouTube-8M [2], provides a feature vector $\mathbf{a} \in \mathbb{R}^{128}$ encoding information contained in the $\mathbf{b}_A$ chunk: $\mathbf{a} = g_A(\mathbf{b}_A; \theta_A)$. Finally, input metadata is normalized according to Table 2, and encoded in $\mathbf{m}_L \in \mathbb{R}^{20}$ and $\mathbf{m}_\mathrm{E} \in \mathbb{R}^{19}$ for *local* and *extended metadata features*, respectively.

Spatiotemporal encodings (STE). Following other transformer-like architectures, we need to add positional encodings to our audiovisual feature embeddings $\mathbf{Z}'$, which can be either learned or fixed. We opt to learn them end-to-end. Being 16 the size of the temporal dimension of the different $\mathbf{Z}'$, we create a vector of zero-centered time indices $\mathbf{t} = \left\langle -\frac{16}{2}, -\frac{16}{2}+1, \ldots, \frac{16}{2}-1 \right\rangle$. The *temporal encodings* are computed by a two-layered network: $\mathbf{P}'_T = \mathrm{ReLU}\left(\Theta_{T_1}^\top \mathrm{ReLU}\left(\Theta_{T_2}^\top \mathbf{t}\right)\right)$, where $\Theta_{T_1} \in \mathbb{R}^{1\times20}$ and $\Theta_{T_2} \in \mathbb{R}^{20\times10}$ are learned weights. The *spatial encodings* $\mathbf{P}'_S$ are computed by a similar encoding network. Given that 28×28 is the spatial resolution of the features, we feed to the spatial encoding network a tensor of spatially zero-centered position indices $\mathbf{S} \in \mathbb{R}^{28\times28\times2}$, where $\mathbf{S}_{i,j} = \left\langle i - \frac{28}{2}, j - \frac{28}{2} \right\rangle, \forall i, j \in [0, 28)$ and weights $\Theta_{S_1} \in \mathbb{R}^{2\times20}$ and $\Theta_{S_2} \in \mathbb{R}^{20\times10}$. Then, $\mathbf{P}'_T$ and $\mathbf{P}'_S$ are reshaped to $\mathbf{P}_T \in \mathbb{R}^{16\times1\times1\times10}$ and $\mathbf{P}_S \in \mathbb{R}^{1\times28\times28\times10}$

Table 2. Description of the different sources of context included as metadata in the proposed personality inference model.

Context type			Source	Value range normalization	Output size
Individual	Stable (across sessions)	Age	Self-reported	$[17, 75] \rightarrow [0, 1]$	1D
		Gender	Self-reported	$\{\mathrm{F}, \mathrm{M}\} \rightarrow \{0, 1\}$	1D
		Cultural background	Self-reported (country of origin)	Recategorization based on cultural differences [53]	6D (one-hot encoding)
	Transient (per session)	Session index	Session info.	$[1, 5] \rightarrow [0, 1]$	1D
		Pre-session mood	Self-reported [32] (8 categories*, Likert scale)	$[1, 5] \rightarrow [0, 1]$ (for each category)	8D
		Pre-session fatigue	Self-reported (Rating scale)	$[0^\dagger, 10] \rightarrow [0, 1]$	1D
Session	Order of the task within the session		Session info.	$[1, 4] \rightarrow [0, 1]$	1D
	Task difficulty†		External survey	$[0, 3] \rightarrow [0, 1]$	1D
Dyadic	Interlocutors' relationship		Self-reported	$\{\mathrm{N}, \mathrm{Y}\} \rightarrow \{0, 1\}$	1D

*Categories: *good*, *bad*, *happy*, *sad*, *friendly*, *unfriendly*, *tense*, and *relaxed*.
† Sessions with fatigue data missing were assigned a value of 0.
‡ Tasks with no difficulty level associated were assigned a value of 0.

and concatenated together by broadcasting singleton dimensions, i.e. $\mathbf{P} = \mathbf{P}_S \parallel \mathbf{P}_T$. $\mathbf{P} \in \mathbb{R}^{16\times28\times28\times20}$ is concatenated to each of the feature embeddings $\mathbf{Z}'$: $\mathbf{Z}_F = \mathbf{Z}'_F \parallel \mathbf{P}$, $\mathbf{Z}_L = \mathbf{Z}'_L \parallel \mathbf{P}$, $\mathbf{Z}_E = \mathbf{Z}'_E \parallel \mathbf{P}$, resulting in $\mathbf{Z}_F, \mathbf{Z}_L, \mathbf{Z}_E \in \mathbb{R}^{16\times28\times28\times148}$. To these features with spatiotemporal encodings, $\mathbf{Z}$, we will later concatenate metadata and audio to obtain the *face query*, *local context*, and *extended context features*.

Query Preprocessor (QP). This small module transforms $\mathbf{Z}_F$ to a vector-form: $\mathbf{f} = \mathrm{QP}(\mathbf{Z}_F)$, $\mathbf{f} \in \mathbb{R}^{128}$. The QP consists of a 3D max pooling layer of size $(1, 2, 2)$ and stride $(1, 2, 2)$, a 3D conv layer of size $(1, 1, 1)$ and 16 filters, a ReLU activation function layer, a permutation of dimensions and reshaping so that the temporal dimensions and the channels are merged into the same dimension, a 2D max pooling of size $(2, 2)$, a 2D conv layer of size $(1, 1)$, a ReLU activation layer, a flattening, and a fully-connected (FC) layer of size 128, another ReLU, and a dropout layer.

Multimodality: fusing visuals with audio and metadata. Both *local* and *extended visual context features* along with encodings, $\mathbf{Z}_L$ and $\mathbf{Z}_E$, are augmented with audio features. The original 128-dimensional global *audio features* $\mathbf{a}$ are linearly projected to a more compact 100-dimensional representation and reshaped to $\mathbf{A} \in \mathbb{R}^{1\times1\times1\times100}$. Then, the *local context features* are simply $\mathbf{W}_L = \mathbf{Z}_L \parallel \mathbf{A}$. The *extended context features* are augmented with the updated audio features and the *extended metadata* from the interlocutor, reshaping $\mathbf{m}_E \in \mathbb{R}^{19}$ to $\mathbf{M}_E \in \mathbb{R}^{1\times1\times1\times19}$ and applying broadcast concatenation, that is $\mathbf{W}_E = \mathbf{Z}_E \parallel \mathbf{A} \parallel \mathbf{M}_E$. Finally, the *face query features* $\mathbf{w}_Q \in \mathbb{R}^{148}$ are built from the combination of the QP output along with the target person's *local metadata*: $\mathbf{w}_Q = \mathbf{f} \parallel \mathbf{m}_L$.

Keys, Values, and Query. To obtain the final input to the transformer layers, we first need to transform *local* and *extended context features* into two different 128-dimensional embeddings (Keys and Values), and also the

Table 3. Evaluated scenarios. Mean value baseline (B) obtained from the mean of the per-trait ground truth labels of the training set; and the proposed method with/without Local (L) and Extended (E) context, Metadata (m), and Audio (a) information.

	Query		**Key** and **Value**			
	Face*	Metadata*	Frame*	Frame‡	Metadata‡	Audio
B	-	-	-	-	-	-
L	✓	-	✓	-	-	-
Lm	✓	✓	✓	-	-	-
LE	✓	-	✓	✓	-	-
LEm	✓	✓	✓	✓	✓	-
LEa	✓	-	✓	✓	-	✓
LEam	✓	✓	✓	✓	✓	✓

* target person and ‡ interlocutor data.

face query features into the query embedding of the same size. The *Local keys* and *Local values* are then $\mathbf{K}_L = \mathrm{ReLU}(\mathbf{\Theta}_{K_L}^\top \mathbf{W}_L)$ and $\mathbf{V}_L = \mathrm{ReLU}(\mathbf{\Theta}_{V_L}^\top \mathbf{W}_L)$ where $\mathbf{\Theta}_{K_L}, \mathbf{\Theta}_{V_L} \in \mathbb{R}^{248\times128}$, whereas the *Extended keys* and *Extended values* are $\mathbf{K}_E = \mathrm{ReLU}(\mathbf{\Theta}_{K_E}^\top \mathbf{W}_E)$ and $\mathbf{V}_E = \mathrm{ReLU}(\mathbf{\Theta}_{V_E}^\top \mathbf{W}_E)$, where $\mathbf{\Theta}_{K_E}, \mathbf{\Theta}_{V_E} \in \mathbb{R}^{267\times128}$. The input *Query* representation $\mathbf{q}_0 \in \mathbb{R}^{128}$ is computed as $\mathbf{q}_0 = \mathrm{ReLU}(\mathbf{\Theta}_{Q_0}^\top \mathbf{w}_Q)$, where $\mathbf{\Theta}_{Q_0} \in \mathbb{R}^{148\times128}$.

Transformer network. Our transformer network (Tx) is composed of $N = 3$ Tx layers with 2 Tx units each, one for the local context and another one for the extended context. The units consist of a multiheaded attention layer with $H = 2$ heads each. Each head computes a separate $128/H$-dimensional linear projection of the query, the keys, and the values, and applies scaled dot product attention as in [75]. Then, it concatenates the H outputs, and linearly projects them back to a new 128-dimensional query. After the multiheaded attention, the resulting query follows the rest of the pipeline in the Tx unit (as illustrated in Fig. 3) to obtain the *updated query*. Note that each unit in the i-th layer provides its own updated query, denoted as $\mathbf{q}_{L_i} \in \mathbb{R}^{128}$ and $\mathbf{q}_{E_i} \in \mathbb{R}^{128}$, $0 < i \leq N$. These are next concatenated together and fed to a FC layer to obtain the i-th layer's joint updated query $\mathbf{q}_i = \mathrm{ReLU}\left(\mathbf{\Theta}_{Q_i}^\top (\mathbf{q}_{\mathrm{L}_i} \parallel \mathbf{q}_{\mathrm{E}_i})\right)$, where $\mathbf{\Theta}_{Q_i} \in \mathbb{R}^{256\times128}$. Finally, $\mathbf{q}_i$ is fed as input to the next ($i+1$-th) layer.

Inference. The per-chunk OCEAN traits are obtained by applying a FC layer to the updated query from the N-th (last) layer, i.e. $\mathbf{y} = \mathbf{\Theta}_{\mathrm{FC}}^\top \mathbf{q}_N$ where $\mathbf{\Theta}_{\mathrm{FC}} \in \mathbb{R}^{128\times5}$. Final per-trait, per-subject predictions are computed as the median of the chunks predictions for each participant.

4.2. Experimental setup

This section describes the experimental setup used to assess the performance of the personality inference model. The evaluation is performed on all tasks except *Gaze*, in which very few personality indicators were present due to the task design. We use frontal camera views (FC1 and FC2, see Fig. 1), in line with the proposed methodology. As personality labels, we use the raw OCEAN scores obtained from the self-reported BFI-2 questionnaire, converted into z-scores using descriptive data from normative samples.

Data and splits description. We use the subset of data composed of participants aged 16 years and above, for which Big-five personality traits are available (see Sec. 3.3). Subject-independent training, validation and test splits were selected following a greedy optimization procedure that aimed at having a similar distribution in each split with respect to participant and session characteristics, while ensuring that no participants appeared in different splits. In terms of sessions and participants, the final splits respectively contain: 116/99 for training, 18/20 for validation, and 11/15 for test. Although the validation split is larger than the test split, the latter contains a better trait balance. Since the duration of the videos is not constant throughout sessions and tasks, in order to balance the number of samples we uniformly selected around 120 chunks from each stream, based on the median number of chunks per video. The final sample of chunks contains 94 960 instances for training, 15 350 for validation and 7 870 for test, distributed among the 4 tasks.

Evaluation protocol. We follow an incremental approach, starting from the *local context*. Six different scenarios are evaluated, summarized in Table 3. We train one model per scenario and task, since each of the four tasks can elicit different social signals and behaviors (detailed in Sec. 3.4), which can be correlated to different degrees with distinct aspects of each personality trait. Results are evaluated with respect to the Mean Squared Error between the aggregated personality trait score and associated ground truth label for each individual in the test set. We also compare the results to a mean value baseline ("B"), computed as the mean of the per-trait ground truth labels of the training set.

4.3. Discussion of results

Obtained per-task results for the different scenarios are shown in Table 4. We discuss some of the findings below.

Effect of including extended (E) visual information. The *extended context* contains visual information from the other interlocutor's behaviors and surrounding scene, allowing the model to consider interpersonal influences during a chunk. By comparing "L" vs. "LE" we can observe that, on average, only *Talk* benefits from the addition of the extended visual context. Trait-wise, *Extraversion* improves for all tasks except for *Lego*, which performs worse for all traits. This can be attributed to the fact that the interaction during this type of collaboration is more slow-paced than in other tasks. Therefore, interpersonal influences cannot be properly captured during just one chunk. In contrast, for more natural tasks such as *Talk*, or fast-moving games such as *Ghost*, there are many instant actions-reactions that can be observed during a single chunk, the effect of which is reflected in the improved results for those tasks. This motivates the need to extend the model to capture longer-time interpersonal dependencies, characteristic of human inter-

Table 4. Obtained results on different tasks. Legend: Mean value baseline (B) obtained from the mean of the per-trait ground truth labels of the training set; and the proposed method with/without Local (L) and/or Extended (E) context, Metadata (m), and Audio (a) information.

	Animals						**Ghost**						**Lego**						**Talk**					
	O	C	E	A	N	**Avg**	O	C	E	A	N	**Avg**	O	C	E	A	N	**Avg**	O	C	E	A	N	**Avg**
B	0.731	0.871	0.988	0.672	1.206	0.894	0.733	0.887	0.991	0.674	1.220	0.901	0.738	0.871	0.990	0.676	1.204	0.896	**0.731**	0.872	0.991	0.673	1.211	0.896
L	0.742	0.879	0.955	0.674	1.133	0.877	0.744	0.891	1.010	0.677	1.242	0.913	**0.723**	0.852	0.917	0.676	1.164	0.866	0.769	0.769	0.997	0.664	1.177	0.875
Lm	**0.721**	0.874	0.946	0.684	1.154	0.876	0.759	0.859	1.027	**0.642**	1.208	0.899	0.725	0.798	0.857	0.618	1.101	0.820	0.743	0.798	0.962	**0.636**	1.168	0.861
LE	0.733	0.832	0.988	0.672	1.221	0.889	0.731	0.905	0.956	0.676	1.291	0.912	0.731	0.885	0.949	0.676	1.230	0.894	0.738	0.793	0.964	0.673	1.094	0.852
LEm	0.736	0.834	0.968	0.669	1.192	0.880	0.743	0.944	0.868	0.657	1.153	0.873	0.727	**0.763**	**0.826**	**0.611**	**1.037**	**0.793**	0.825	**0.718**	0.878	0.639	1.047	0.821
LEa	0.722	0.827	0.954	0.672	1.211	0.877	**0.730**	**0.872**	0.950	0.672	1.199	0.885	0.742	0.867	0.941	0.672	1.229	0.890	0.757	0.728	0.970	0.664	1.106	0.845
LEam	0.737	**0.756**	**0.887**	**0.580**	**1.023**	**0.797**	0.741	0.893	**0.844**	0.667	**1.139**	**0.857**	0.745	0.839	0.953	0.659	1.099	0.859	0.773	0.790	**0.869**	0.670	**0.985**	**0.817**

actions, across a series of ordered chunks along time, to truly benefit from this extended information.

Effect of including metadata (m) information. The inclusion of metadata validates our intuition that personal, task, and dyadic details provide relevant information to the model to produce overall better predictions, particularly if the cases "L" vs. "Lm", "LE" vs. "LEm", and "LEa" vs. "LEam" are compared, with the largest improvement observed for *Lego* (11.29%, "LE" vs. "LEm" case). Considering the high heterogeneity and dimensionality of behaviors revealed in an interaction and their multiple meanings, these concise features appear to be beneficial to better guide the model and establish meaningful patterns in the data. Nonetheless, a systematic study would be needed to assess the effect of each feature individually.

Effect of including audio (a) information. From comparing "LE" vs. "LEa" and "LEm" vs. "LEam", we observe that better results are obtained, on average, for all the tasks when audio information is considered. In line with previous literature [77], it is clear that paralinguistic acoustic features are required to better model personality. However, the observed improvement is smaller for *Lego*. One plausible reason would be the noise produced by the Lego pieces while being moved, or by the instructions book while turning its pages close to the microphones, which would interfere with the learning process. In the case of more natural routines like *Talk*, the influence of audio is not as strong as we would have expected. In contrast, *Animals*, another speaking-based task, obtains the best results for almost all traits when audio is considered. There is one salient difference among these two tasks that may explain this pattern. The latter elicits more individual covert thinking and cognitive processes that cannot be entirely observed from the visual modality, so most of the overt information comes from the spoken conversation. In contrast, the former elicits a larger range of visual cues which may be more relevant than acoustic features for certain traits.

Putting everything together. In the last experiment ("LEam"), the model is aware of the overall contextual information. We notice that apart from *Lego*, for which the audio drawbacks were already commented, all the other tasks seem to benefit from the provided knowledge, obtaining the lowest error value on average.

Baseline comparison. We observe that *Agreeableness*, followed by *Openness*, obtain the lowest error among mean value baseline ("B") results, indicating that ground truth labels for such traits are more concentrated. In those cases, none of the models achieve a substantial improvement over the baseline, except for *Animals*, where "LEam" obtains an error of 0.58, the lowest overall. At the other end we find *Neuroticism*, which is the trait with most spread values, but also the one for which we obtain the largest benefits with the evaluated models. In particular, the largest improvement overall (18.66%) is given by "LEam" for *Talk*.

5. Conclusion

This paper introduced UDIVA, the largest multiview audiovisual dataset of dyadic face-to-face non-scripted interactions. To validate part of its potential, we proposed a multimodal transformer-based method for inferring the personality of a target person in a dyadic scenario. We incrementally combined different sources of context (both interlocutors' scene, acoustic and task information) finding consistent improvements as they were added, which is consonant with human interaction research in the psychology field.

UDIVA is currently being annotated with additional labels (e.g., transcriptions, continuous action/intention for human-object-human interaction) to allow for a more holistic analysis of human interaction from both individual and dyadic perspectives. From a methodological point of view, we plan to extend the proposed architecture to better capture long-term discriminative features. Nevertheless, we are releasing this data[5] with the purpose of advancing the research and understanding of human communication from a multidisciplinary perspective, far beyond personality analysis.

Acknowledgements. This work has been partially supported by the Spanish projects TIN2015-66951-C2-2-R, RTI2018-095232-B-C22, PID2019-105093GB-I00 (MINECO/FEDER, UE), CERCA Programme/Generalitat de Catalunya, and ICREA under the ICREA Academia programme. We gratefully thank Alejandro Alfonso, Rubén Ballester, Rubén Barco, Aleix Casellas, Roger Hernando, Andreu Masdeu, Carla Morral, Pablo Lázaro, Luis M. Pérez, Italo Vidal, Daniel Vilanova, all HUPBA members, and all participants for their support in the dataset collection phase. We also thank the Department of Mathematics and Informatics of Universitat de Barcelona for granting facilities access, the Data Protection and Bioethics departments for their assistance, and Noldus IT for loaning part of the recording setup.

[5]The dataset will be available at `http://chalearnlap.cvc.uab.es/dataset/39/description/`.

References

[1] Jennifer Abel and Molly Babel. Cognitive load reduces perceived linguistic convergence between dyads. *Language and Speech*, 60(3):479–502, 2017.

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016.

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

[4] Icek Ajzen. Attitudes, traits, and actions: Dispositional prediction of behavior in personality and social psychology. In *Advances in experimental social psychology*, volume 20, pages 1–63. Elsevier, 1987.

[5] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1707–1720, 2015.

[6] Oya Aran and Daniel Gatica-Perez. Cross-domain personality prediction: from video blogs to small group meetings. In *International Conference on Multimodal Interaction*, pages 127–130, 2013.

[7] Damaris Aschwanden, Mathias Allemand, and Patrick L Hill. Cognitive methods in personality research. *The Wiley Encyclopedia of Personality and Individual Differences: Measurement and Assessment*, pages 49–54, 2020.

[8] Michael Ashton and Kibeom Lee. The hexaco-60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91:340–5, 07 2009.

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

[10] Sanjay Bilakhia, Stavros Petridis, Anton Nijholt, and Maja Pantic. The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61, 2015.

[11] Philippe Blache, Roxane Bertrand, Gaëlle Ferré, Berthille Pallaud, Laurent Prévot, and Stéphane Rauzy. The corpus of interactional data: A large multimodal annotated resource. In *Handbook of linguistic annotation*, pages 1323–1356. Springer, 2017.

[12] Dawn O Braithwaite and Paul Schrodt. *Engaging theories in interpersonal communication: Multiple perspectives*. Sage Publications, 2014.

[13] Judee K Burgoon, Lesa A Stern, and Leesa Dillman. *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press, 2007.

[14] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335, Nov 2008.

[15] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, January-March 2017.

[16] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. The noxi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 350–359, 2017.

[17] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement. *IEEE Transactions on Affective Computing*, 2017.

[18] Huili Chen, Yue Zhang, Felix Weninger, Rosalind Picard, Cynthia Breazeal, and Hae Won Park. Dyadic speech-based affect recognition using dami-p2c parent-child multimodal interaction dataset. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 97–106, 2020.

[19] Herbert H Clark, Robert Schreuder, and Samuel Buttrick. Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, 22(2):245–258, 1983.

[20] Gerald L Clore and Jeffrey R Huntsinger. How emotions inform judgment and regulate thought. *Trends in cognitive sciences*, 11(9):393–399, 2007.

[21] 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.

[22] 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.

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

[24] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, et al. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *International conference on affective computing and intelligent interaction*, pages 488–500. Springer, 2007.

[25] Ellen Douglas-Cowie, Cate Cox, Jean-Claude Martin, Laurence Devillers, Roddy Cowie, Ian Sneddon, Margaret McRorie, Catherine Pelachaud, Christopher Peters, Orla Lowry, et al. The humaine database. In *Emotion-Oriented Systems*, pages 243–284. Springer, 2011.

[26] 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.

[27] Rana El Kaliouby, Rosalind Picard, and Simon Baron-Cohen. Affective computing and autism. *Annals of the New York Academy of Sciences*, 1093(1):228–248, 2006.

[28] Lesa Ellis and Mary Rothbart. Revision of the early adolescent temperament questionnaire. *Poster presented at the 2001 Biennal Meeting of the Society for Research in Child Development*, 01 2001.

[29] Anna Esposito and Lakhmi C Jain. Modeling emotions in robotic socially believable behaving systems. In *Toward Robotic Socially Believable Behaving Systems-Volume I*, pages 9–14. Springer, 2016.

[30] Sheng Fang, Catherine Achard, and Séverine Dubuisson. Personality classification and behaviour interpretation: An approach based on feature categories. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 225–232, 2016.

[31] Riccardo Fusaroli, Johanne S Bjørndahl, Andreas Roepstorff, and Kristian Tylén. A heart for interaction: Shared physiological dynamics and behavioral coordination in a collective, creative construction task. *Journal of Experimental Psychology: Human Perception and Performance*, 42(9):1297, 2016.

[32] David Gallardo-Pujol, Antonio Andres-Pueyo, and Alberto Maydeu-Olivares. Maoa genotype, social exclusion and aggression: An experimental test of a gene-environment interaction. *Genes, brain, and behavior*, 12, 10 2012.

[33] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12046–12055, 2019.

[34] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[35] Willard W Hartup, Doran C French, Brett Laursen, Mary Kathleen Johnston, and John R Ogawa. Conflict and friendship relations in middle childhood: Behavior in a closed-field situation. *Child Development*, 64(2):445–454, 1993.

[36] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2122–2132, 2018.

[37] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[38] 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, 2017.

[39] J. C. S. Jacques Junior, Y. Güçlütürk, M. Perez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. Van Gerven, R. Van Lier, and S. Escalera. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, pages 1–1, 2019.

[40] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[41] David A Kenny. Models of non-independence in dyadic research. *Journal of Social and Personal Relationships*, 13(2):279–294, 1996.

[42] Michael Kipp, Jean-Claude Martin, Patrizia Paggio, and Dirk Heylen. *Multimodal corpora: from models of natural interaction to systems and applications*, volume 5509. Springer, 2009.

[43] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjoern W Schuller, et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[44] Chi-Chun Lee, Carlos Busso, Sungbok Lee, and Shrikanth S Narayanan. Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[45] Jeng-Lin Li and Chi-Chun Lee. Attention learning with retrievable acoustic embedding of personality for emotion recognition. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 171–177. IEEE, 2019.

[46] Qingbiao Li, Chunhua Wu, Zhe Wang, and Kangfeng Zheng. Hierarchical transformer network for utterance-level emotion recognition. *Applied Sciences*, 10(13):4447, 2020.

[47] Zheng Lian, Jianhua Tao, Bin Liu, Jian Huang, Zhanlei Yang, and Rongjun Li. Context-dependent domain adversarial neural network for multimodal emotion recognition. *Proc. Interspeech 2020*, pages 394–398, 2020.

[48] 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.

[49] Sally Lindsay, Kara Grace Hounsell, and Celia Cassiani. A scoping review of the role of lego® therapy for improving inclusion and social skills among children and youth with autism. *Disability and health journal*, 10(2):173–182, 2017.

[50] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825, 2019.

[51] 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.

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

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

[54] Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152, 2013.

[55] Angeliki Metallinou, Zhaojun Yang, Chi-Chun Lee, Carlos Busso, Sharon Carnicke, and Shrikanth Narayanan. The USC creativeit database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations. *Lang. Resour. Eval.*, 50(3):497–521, 2016.

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

[57] Iftekhar Naim, M Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6, 2015.

[58] Shrikanth Narayanan and Panayiotis G Georgiou. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, 101(5):1203–1233, 2013.

[59] 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.

[60] Nuria Osa, Roser Granero, Eva Penelo, Josep Domènech, and Lourdes Ezpeleta. The short and very short forms of the children's behavior questionnaire in a community sample of preschoolers. *Assessment*, 21, 11 2013.

[61] Hye Jeong Park and Jae Hwa Lee. Looking into the personality traits to enhance empathy ability: A review of literature. In *International Conference on Human-Computer Interaction*, pages 173–180, 2020.

[62] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[63] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

[64] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953, 2019.

[65] Ricardo Darío Pérez Principi, Cristina Palmero, Julio C Junior, and Sergio Escalera. On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, 2019.

[66] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, 2020.

[67] James Rehg, Gregory Abowd, Agata Rozga, Mario Romero, Mark Clements, Stan Sclaroff, Irfan Essa, O Ousley, Yin Li, Chanho Kim, et al. Decoding children's social behavior. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3414–3421, 2013.

[68] Mary Rothbart, Stephan Ahadi, Karen Hershey, and Philip Fisher. Investigations of temperament at three to seven years: The children's behavior questionnaire. *Child development*, 72:1394–408, 09 2001.

[69] Ullica Segerstrale and Peter Molnár. *Nonverbal communication: where nature meets culture*. Routledge, 2018.

[70] 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.

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

[72] 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.

[73] 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.

[74] Marion K Underwood, Bertrina L Scott, Mikal B Galperin, Gretchen J Bjornstad, and Alicia M Sexton. An observational study of social exclusion under varied conditions: Gender and developmental differences. *Child Development*, 75(5):1538–1555, 2004.

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. 2017.

[76] 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.

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

[78] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743 – 1759, 2009.

[79] 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.

[80] 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.

[81] 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.

[82] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[83] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *AAAI Conference on Artificial Intelligence*, volume 2018, page 5642, 2018.

[84] Aurélie Zara, Valérie Maffiolo, Jean Claude Martin, and Laurence Devillers. Collection and annotation of a corpus of human-human multimodal interactions: Emotion and others anthropomorphic characteristics. In *International Conference on Affective Computing and Intelligent Interaction*, pages 464–475. Springer, 2007.

[85] 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, 2020.

[86] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2018.

[87] 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, 2016.

[88] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176, 2019.