– Supplementary Material –
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

Here we provide further information on some sections of the paper. First, we include an extended table containing a more complete comparison of available dyadic interaction datasets. Then, we describe the rationale behind using 32 frames per video chunk and the procedure used to crop the face-only videos, as part of the proposed methodology. Finally, we detail the training strategy, the algorithm used to define the data splits (such that a balance was kept on the participants and sessions features) and report the resulting distribution of OCEAN values among them.

S1. Face-to-face dyadic datasets comparison (Sec. 2)

For the sake of completeness, Table S1 contains an extended review of publicly available face-to-face dyadic interactions datasets that contain at least audiovisual data. Most of the datasets are tailor-made for too specific purposes or limited in the number of participants, recordings, views, context annotations or language. Hence, there is no big enough general purpose database in the literature that could allow for an integral analysis of both, the interaction and the participants.

S2. Size of video chunks (Sec. 4.1)

The original Video Action Transformer [13] uses an I3D backbone pretrained on Kinetics-400 [6] for spatiotemporal feature extraction. Such backbone uses 64 frames per chunk, which is equivalent to around 3 seconds of video. Instead, we opted for the R(2+1)D backbone [25] pretrained on IG-65M dataset, which has shown to provide significant performance gains [12]. This backbone uses 32 frames per chunk, so by using a stride of 2 we manage to encode approximately the same time window as the original method with half the number of frames while reducing the memory load. This is equivalent to downsampling the original videos from 25 fps to 12.5, that is, 1 frame every 0.08 seconds. Although not frequent, there is a chance to miss some fast-paced facial and body micro-actions in such downsampling process. However, there is also the trade-off we try to balance between losing some of these fast micro-actions and being able to include a larger, and also important, temporal context.

S3. Face detection and tracking (Sec. 4.1)

As described in the main paper, we use a face chunk video as one of the inputs of the model, which is used together with the participants' metadata to form the query of the transformer model. In order to detect the faces we use MobileNet-SSD [14], deployed using Tensorflow Object Detection API [15] and pretrained on the Wider Face Dataset [27]. As we consider only frontal cameras, the detection task is not very challenging, therefore, on more than 95% of the videos the detection ratio is higher than 75%. In case the gap between consecutive detections is lower than 25 frames (1 second), we linearly interpolate the coordinates of the boxes. Since there are frames in which the frontal cameras capture both participants, we need to identify the target person before computing the face chunks. In order to do so, we employ a basic tracking algorithm based on the following 2 steps: (1) *identify* target person's face: given a video, the face of the target person is considered the first detection that has a mean intersection over union (IoU) score higher than 0.2 with respect to all the other faces in the video; (2) *track* target person face throughout the video based on the IoU.

*These authors contributed equally to this work.

Table S1. 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.
IFADV [24] ('07)	Speech & conversation analysis	Non-acted	Audiovisual	Speech features, transcripts	24/10	20	5h	2	Dutch
IEMOCAP [4] ('08)	Emotion recognition	Acted* & Acted	Audiovisual, face & hands MoCap.	Emotions, transcripts, turn-taking	5/5	5	$\sim$12h	2	English
CID [2] ('08)	Speech & conversation analysis	Non-acted & Non-acted*	Audiovisual	Speech features, transcripts	10/6	8	8h	1	French
Spontal [11] ('10)	Speech & conversation analysis	Non-acted & Non-acted*	Audiovisual, head & torso MoCap.	Transcripts, speech features	ϕ	120	60h	2	Swedish
NOMCO [21] ('10)	Speech & conversation analysis	Non-acted & Non-acted*	Audiovisual	Speech & interaction features, gestures, transcripts, emotions	6/6 $^{\phi}$	60	$\sim$6h	3	Danish, Swedish, Finnish
HUMAINE† [9, 10] ('11)	Emotion analysis	Non-acted*	Audiovisual	Emotions	34	18	$\sim$12h	4	English
MMDB [23] ('13)	Adult-infant interaction analysis	Non-acted*	Audiovisual, depth, physiological	Social cues (gaze, vocal affects, gestures...)	121	160	$\sim$13.3h	8 + 1D	English
MAMCO [26] ('14)	Overlap analysis	Non-acted	Audiovisual	Transcripts	6/6	12	$\sim$1h	3	Maltese
4D CCDb [16] ('15)	Speech & conversation analysis	Non-acted	Audiovisual, depth	Facial expressions, head gestures, utterances	2/2	17	$\sim$0.2h	6 + 8M	English
MAHNOB [1] ('15)	Mimicry	Non-acted*	Audiovisual, head MoCap.	Head, face and hand gestures, personality scores (self-reported)	29/31	54	11.6h	2 + 13M	English
MIT Interview [20] ('15)	Hirability analysis	Non-acted*	Audiovisual	Hirability, speech features, social & behavioral traits, transcripts	43/26	138	10.5h	2	English
MPIIEMO [19] ('15)	Bodily emotion analysis	Acted	Audiovisual	Emotions	3/2	8$\times$7$\times$4 (tasks)	$\sim$2.4h	8	German$^{\phi}$
JESTKOD [3] ('15)	Agreement classification	Non-acted*	Audiovisual, body MoCap.	Agreement, emotion	4/6	25	4.3h	1	Turkish
Creative IT [18] ('16)	Emotion recognition	Acted	Audiovisual, body MoCap.	Transcripts, speech features, emotion	9/7	8	$\sim$1h	2	English
MSP-IMPROV [5] ('17)	Emotion recognition	Acted & Non-acted	Audiovisual	Turn-taking, emotion	6/6	6	9h	2	English
NNIME [8] ('17)	Emotion analysis	Non-acted*	Audiovisual, physiological	Emotion, transcripts	22/20	102	$\sim$11h	1	Chinese
RAMAS [22] ('18)	Emotion analysis	Non-acted* & Acted	Audiovisual, depth, body MoCap.	Physiological signals, emotion, interaction traits	5/5	80	$\sim$7h	2 + 1D	Russian
DAMI-P2C [7] ('20)	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) ('20)	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$\times$5 (tasks)	90.5h	6 + 2E	Spanish, Catalan, English

† Here we consider the Green Persuasive and the EmoTABOO [28] datasets together.

S4. Training strategy (Sec. 4.2)

The proposed model was trained using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$ and a learning rate of $1e-5$. We used a batch size of 2 and the Mean Squared Error as the loss function. We compute the validation error approximately 30 times per epoch and select the model that gives the best results considering the mean with its previous and next evaluation scores. The final results, detailed in Sec. 4.3 of the main paper, were obtained by freezing the layers of the R(2+1)D backbones, as strategies such as finetuning end-to-end or only the last block of the feature extractors led to fast overfitting.

S5. Personality trait (OCEAN) values over splits (Sec. 4.2)

In this section, we briefly describe the procedure used to define the data splits used during the experiments described in the experimental section.

In order to split the data among training, validation and test subsets, some sessions needed to be removed so that no participants were repeated in any of the subsets. The final split was selected using a greedy optimization method that iteratively removed and added sessions based on their importance until a valid split ratio was found. Such importance was determined by the groups distribution and the number of remaining sessions per participant. In particular, the method tried to minimize a set of costs to: (1) ensure that distributions among splits were not different by means of a Kolmogorov-Smirnov significance test [17]; (2) ensure that Pearson's correlation of gender, age and personality values among splits did not differ by a large margin; (3) attempt to have a uniform distribution in validation and test with respect to age and gender to correct selection bias; (4) attempt to have a close-to-uniform distribution of group combinations; and (5) try to maximize the number of sessions without losing participants, while considering also the train/validation/test ratio. The resulting distribution of OCEAN values among splits can be seen in Fig. S1.

References

[1] 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.

[2] 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.

[3] Elif Bozkurt, Hossein Khaki, Sinan Keçeci, B Berker Türker, Yücel Yemez, and Engin Erzin. Jestkod database: Dyadic interaction analysis. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pages 1374–1377. IEEE, 2015.

[4] 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.

[5] 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.

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[7] 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.

[8] Huang-Cheng Chou, Wei-Cheng Lin, Lien-Chiang Chang, Chyi-Chang Li, Hsi-Pin Ma, and Chi-Chun Lee. Nnime: The nthu-ntua chinese interactive multimodal emotion corpus. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 292–298. IEEE, 2017.

[9] 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.

[10] 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.

[11] Jens Edlund, Jonas Beskow, Kjell Elenius, Kahl Hellmer, Sofia Strömbergsson, and David House. Spontal: A swedish spontaneous dialogue corpus of audio, video and motion capture. In *LREC*, pages 2992–2995, 2010.

[12] 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.

[13] 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.

[14] 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.

[15] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3296–3297, 2017.

[16] Andrew David Marshall, Paul L Rosin, Jason Vandeventer, and Andrew Aubrey. 4d cardiff conversation database (4d ccdb): A 4d database of natural, dyadic conversations. *Auditory-Visual Speech Processing,{AVSP} 2015*, pages 157–162, 2015.

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

[18] 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.

[19] Philipp M Müller, Sikandar Amin, Prateek Verma, Mykhaylo Andriluka, and Andreas Bulling. Emotion recognition from embedded bodily expressions and speech during dyadic interactions. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 663–669. IEEE, 2015.

[20] 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.

[21] P Paggio, J Allwood, Jokinen Ahlsén, and K Jokinen. The nomco multimodal nordic resource-goals and characteristics. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 10) Valletta,*

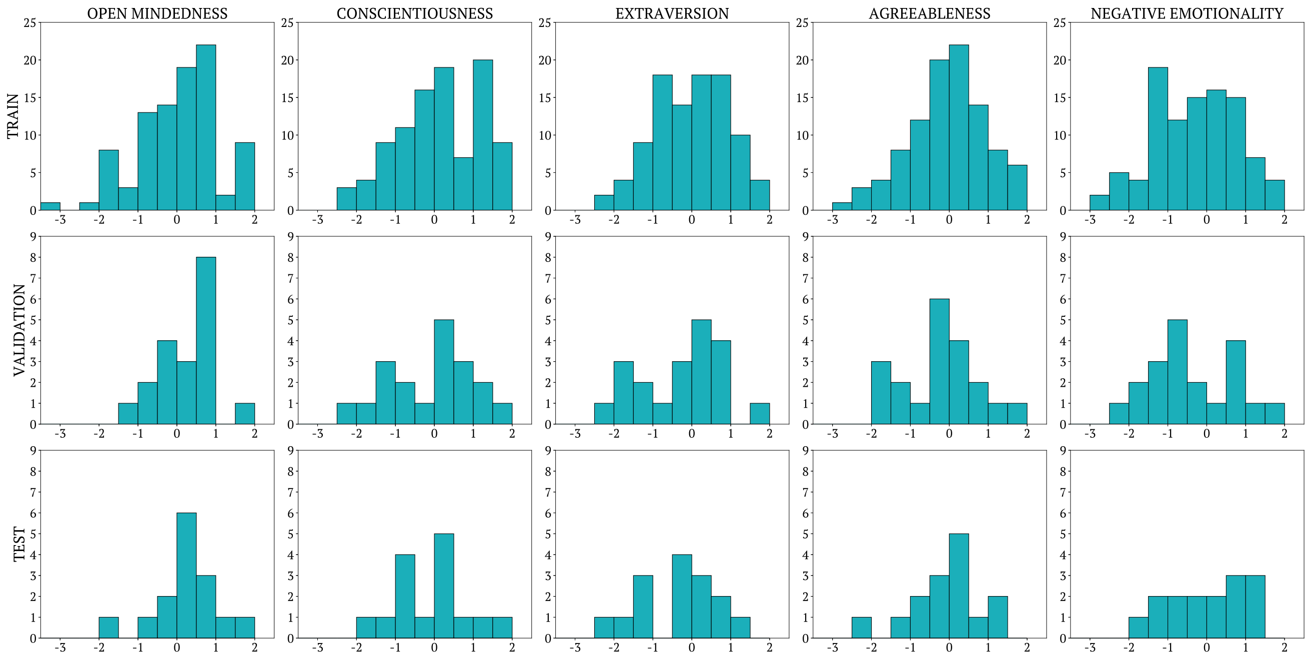

Figure S1. Distribution of the self-reported personality trait (OCEAN) values across train, validation and test splits used to evaluate the proposed personality inference method. X axis refers to z scores for each personality trait. Y axis refers to number of participants.

Malta. May, pages 19–21. European Language Resources Association (ELRA), 2010.

[22] Olga Perepelkina, Evdokia Kazimirova, and Maria Konstantinova. Ramas: Russian multimodal corpus of dyadic interaction for affective computing. In *International Conference on Speech and Computer*, pages 501–510. Springer, 2018.

[23] 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.

[24] Eric Sanders Rob van Son, Wieneke Wesseling and Henk van den Heuvel. The ifadv corpus: a free dialog video corpus. In Bente Maegaard Joseph Mariani Jan Odijk Stelios Piperidis Daniel Tapias Nicoletta Calzolari (Conference Chair), Khalid Choukri, editor, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may 2008. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[25] 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.

[26] Alexandra Vella and Patrizia Paggio. Overlaps in maltese: a comparison between map task dialogues and multimodal conversational data. In *NEALT Proceedings. Northern European Association for Language and Technology; 4th Nordic Symposium on Multimodal Communication; November 15-16; Gothenburg; Sweden*, number 093, pages 21–29. Linköping University Electronic Press, 2013.

[27] 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.

[28] 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.