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

Supplementary Material

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

Appendix A. UDIVA v0.5 split statistics

This section includes additional graphs and tables depicting the differences in data splits associated to Section 2.3 of the main paper. In particular, we show differences with respect to pre- and post-session fatigue (Figure 1) and mood (Figure 2), as well as gender (Table 1 and Figure 3) and age (Table 2 and Figure 4) differences with respect to personality traits.

Appendix B. Hands post-processing

The interlocutors' hands appear very frequently in the views used for this work (*FC1* and *FC2*), as can be observed in Figure 5 of the main paper (Section 2.4.4). In such scenarios, landmark extraction methods may generate landmarks for more than one person. We easily retrieved the face and body of interest by selecting the most centered set of landmarks. Unfortunately, the highly sparse set of locations where the hands of interest appear overlaps with those from the confounding hands (the interlocutor's). This sets off a high number of frames where the hand of interest for a given interlocutor can be easily mistakenly chosen. The same applies to the right-left association of the hands, which might be wrong if the arms of the person are crossed. By default, the hands estimator module (Rong et al., 2021) associates the left/right label if the detection is the closest to the left/right body elbow (which is inferred in parallel). This assumption becomes especially erroneous for the aforementioned crossed-arms scenario.

In order to minimize these errors, we sequentially applied several post-processing methods:

1. We extracted the hands bounding boxes and their left/right default association.

2. We removed temporally consecutive detections with an Intersection-over-Union (IoU) value smaller than 0.1. This step mainly helped to discard hand switches or false detections.

3. Missed and removed left/right hand detections increased the number of segments of consecutive frames where no left/right detections were available. These hands gaps were tracked forwards and backwards using the method from Li et al. (2019). If the bidirectional tracker eventually overlapped with the last and first detections before and after the gap (i.e., tracked detection has an IoU with the original detection bigger than 0.25), both sequences of tracked detections from the gap were merged and saved. If only one tracker overlapped with the other side of the gap, the merging was skipped and only the detections from the successful tracker were stored. This step successfully recovered all the correct detections removed in the previous step and fixed sequences of rapidly moving hands.

4. We extracted the hands landmarks inside the bounding boxes from the previous step.

5. We used the x and y coordinates of the body pose landmarks from the person of interest to compute a 2D vector from the left wrist to the left palm, V_{LB}. Using the left hand landmarks, we extracted the 2D vector from the left wrist to the left middle knuckle, V_{LH}. Similarly, we computed V_{RB} and V_{RH} for the right hand. If

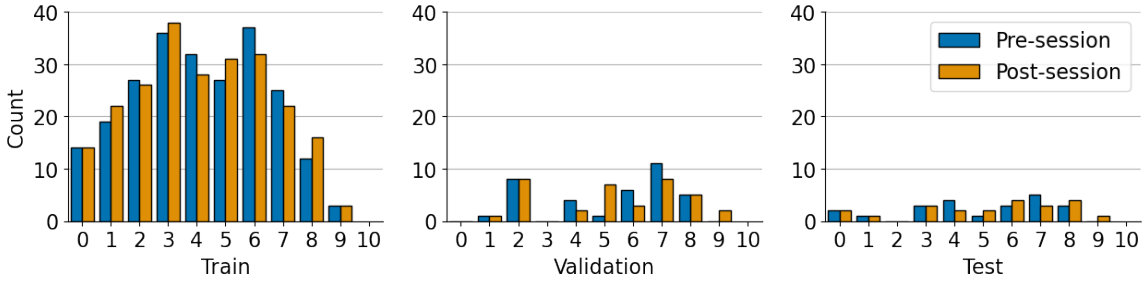

Figure 1: Pre- and post-session fatigue distribution across train, validation and test splits of the UDIVA v0.5 dataset.

Trait	Training	Validation	Test
O	t(97)=-0.07; p=0.94; d=0.01	t(18)=0.62; p=0.54; d=0.3	t(13)=0.85; p=0.41; d=0.47
C	t(97)=2.45; p=0.02; d=0.5	t(18)=0.6; p=0.56; d=0.28	t(13)=2.64; p=0.02; d=1.47
E	t(97)=1.78; p=0.08; d=0.36	t(18)=0.64; p=0.53; d=0.3	t(13)=1.64; p=0.13; d=0.91
A	t(97)=2.65; p=0.01; d=0.54	t(18)=-0.32; p=0.75; d=0.15	t(13)=2.13; p=0.05; d=1.18
N	t(97)=2.71; p=0.01; d=0.55	t(18)=1.09; p=0.29; d=0.52	t(13)=1.65; p=0.12; d=0.92

Table 1: Gender differences in OCEAN scores on training, validation and test splits of the UDIVA v0.5 dataset, by means of Student's t-test.

$V_{LB} \cdot V_{LH} < 0$ and $V_{RB} \cdot V_{RH} < 0$, i.e., the angles between the hand orientation according to the hands landmarks and the hand orientation according to the body pose landmarks differed more than 90° for both hands, the left/right associations were switched.

6. We removed the hands detections that fulfilled any of the following conditions:
   - Left/right hand wrist was closer to the left/right body pose wrist of the confounder (i.e., interlocutor) by a margin of 20 pixels.
   - The left hand wrist was closer to the right body pose wrist by a margin of 60 pixels with respect to the left body pose wrist. Similarly for the right hand wrist.
7. We repeated the steps 3 and 4, filling the gaps generated by the previous step, and extracting the new and final hand landmarks.
8. We applied the one-euro filter to the hand landmarks with a cut-off of 0.001 and a $\beta = 0.02$ (Casiez et al., 2012).

Some of the previous parameters were selected after analyzing the properties of the resulting landmarks (e.g., distances between hands and wrists, angles, etc.) and others by visually inspecting the erroneous results (e.g., intersection-over-union thresholds).

Trait	Training	Validation	Test
O	-0.19 p=0.057 [-0.37;0.01]	-0.34 p=0.146 [-0.68;0.12]	0.01 p=0.981 [-0.51;0.52]
C	0.35 p<0.001 [0.17;0.51]	0.11 p=0.632 [-0.35;0.53]	0.38 p=0.165 [-0.17;0.75]
E	-0.01 p=0.924 [-0.21;0.19]	0.21 p=0.364 [-0.25;0.6]	0.09 p=0.742 [-0.44;0.58]
A	0.26 p=0.01 [0.07;0.43]	0.33 p=0.154 [-0.13;0.67]	-0.21 p=0.46 [-0.65;0.34]
N	-0.13 p=0.218 [-0.31;0.07]	-0.06 p=0.795 [-0.49;0.39]	-0.36 p=0.185 [-0.74;0.18]

Table 2: Statistical tests and 95% CIs for the correlations between OCEAN scores and age of the UDIVA v.05 splits.

Appendix C. Behavior forecasting: segments distribution

The distribution of candidate and final segments after clustering them as described in Section 5.3.2 of the main paper are shown in Figure 5. Representative behaviors from segments belonging to four different clusters are shown in Figure 6.

References

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.

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.

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.

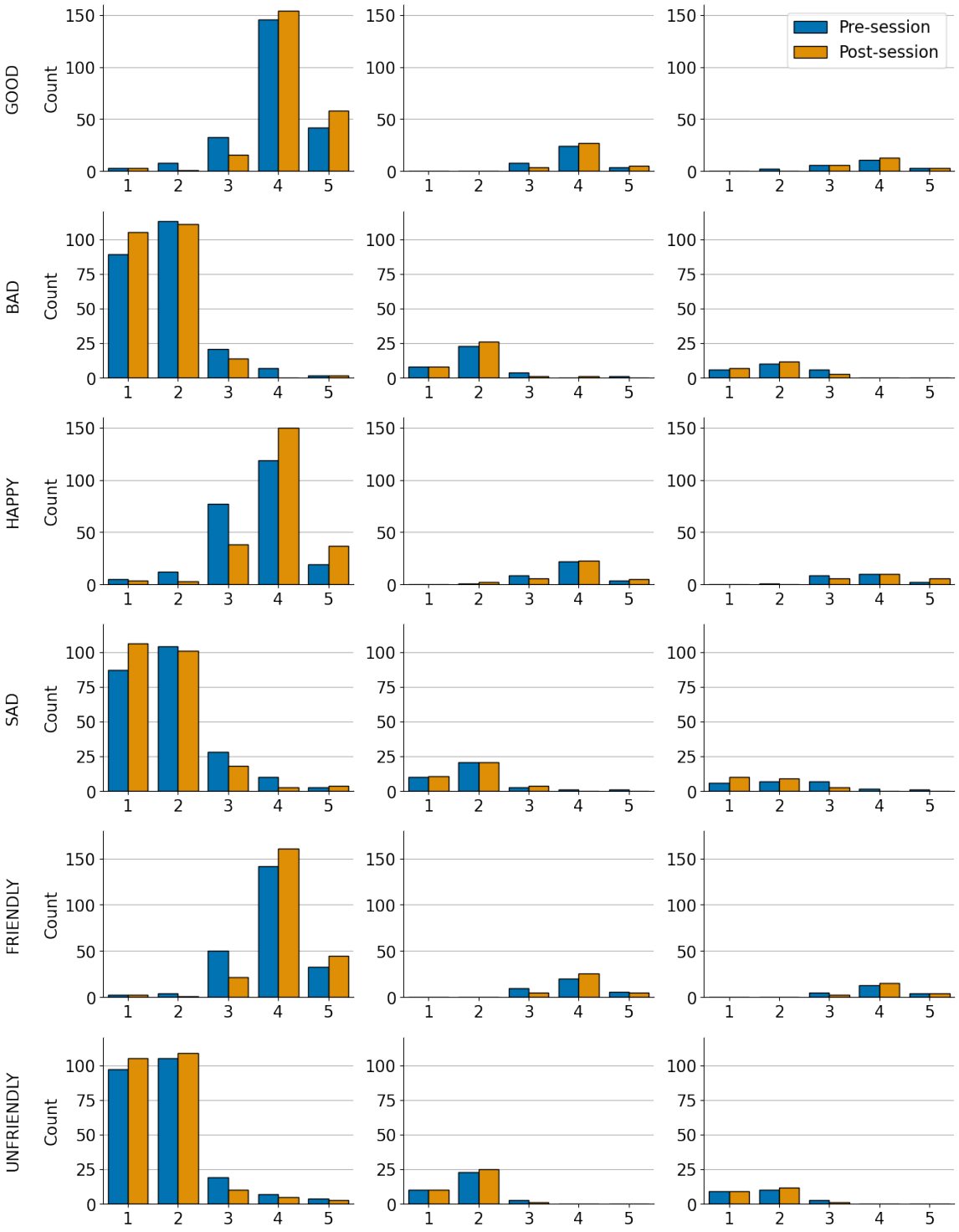

Figure 2: Pre- and post-session distribution of mood categories across train, validation and test splits of the UDIVA v0.5 dataset.

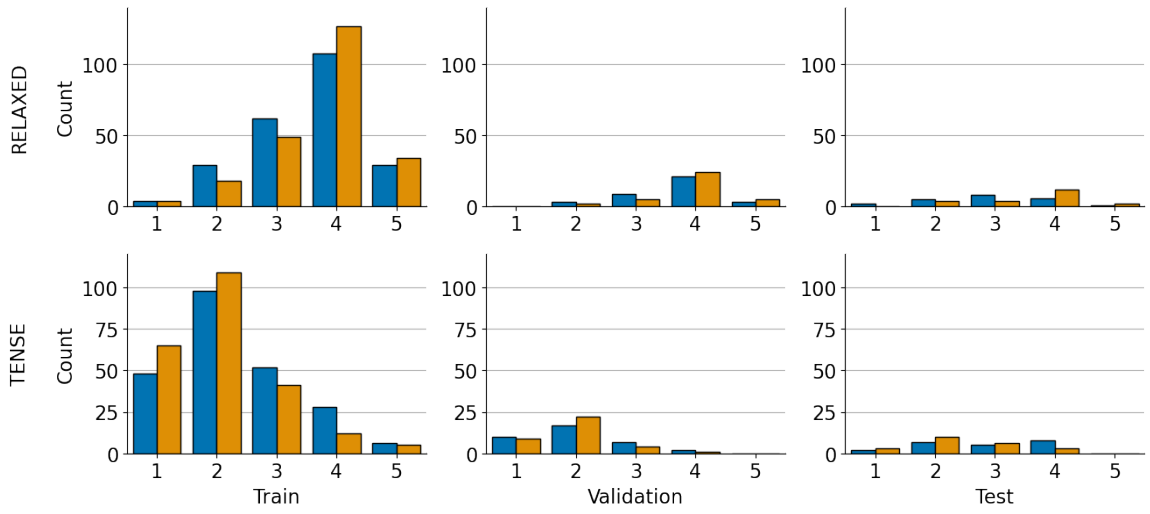

Figure 2: (Continuation) Pre- and post-session distribution of mood categories across train, validation and test splits of the UDIVA v0.5 dataset.

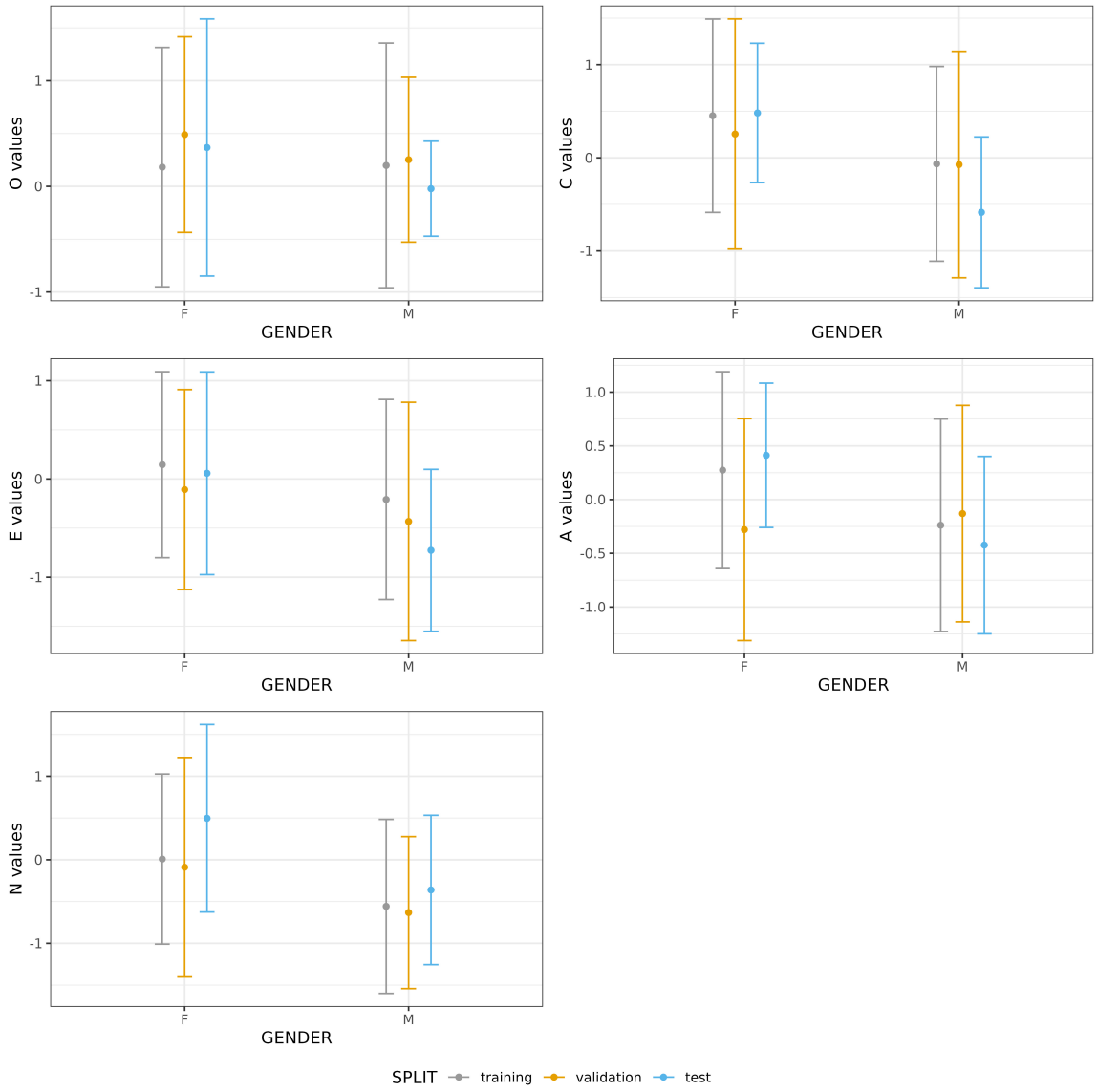

Figure 3: Gender differences in OCEAN scores on training, validation and test splits of the UDIVA v0.5 dataset.

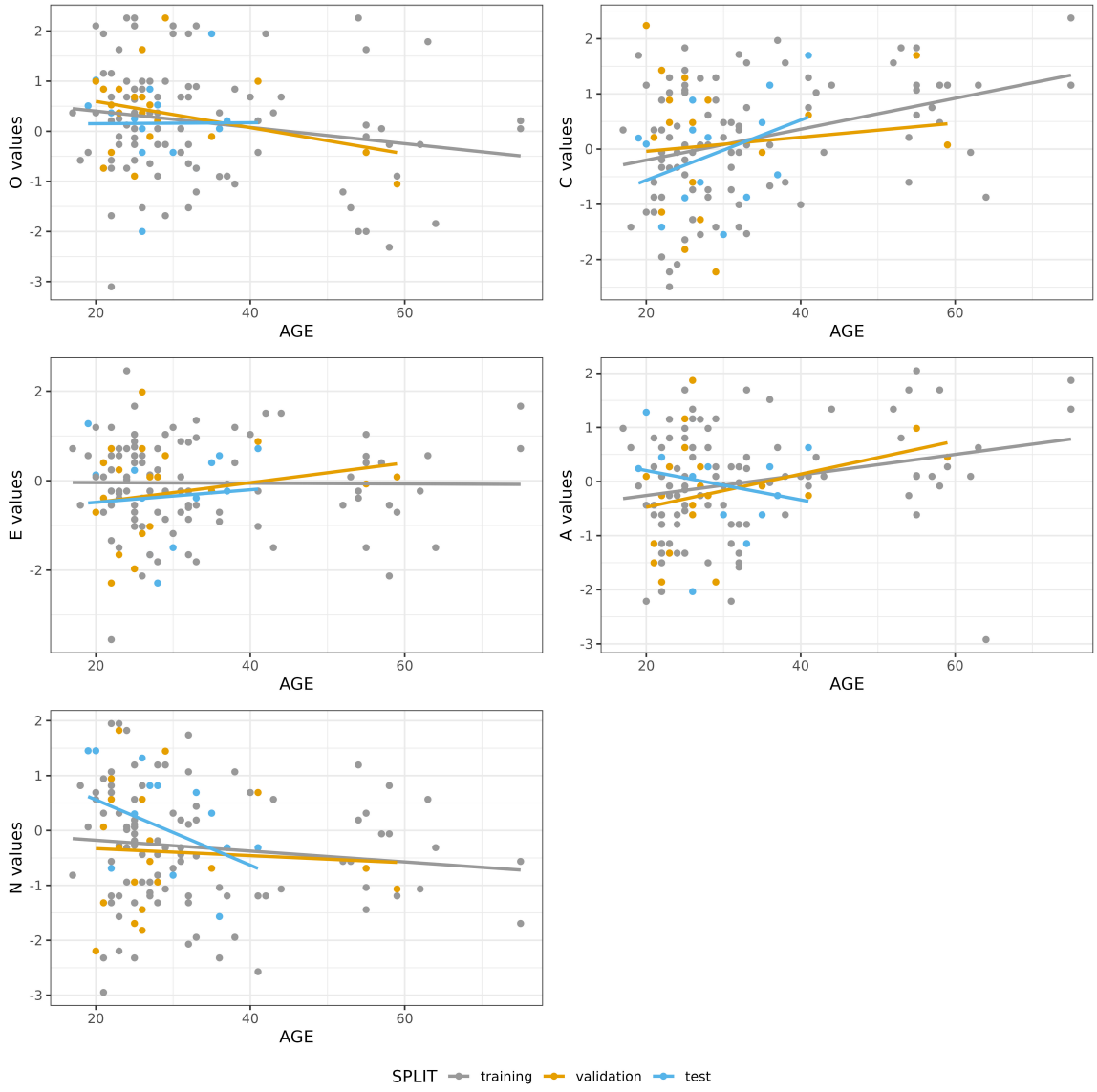

Figure 4: Relationship between OCEAN scores and age along training, validation and test splits of the UDIVA v0.5 dataset.

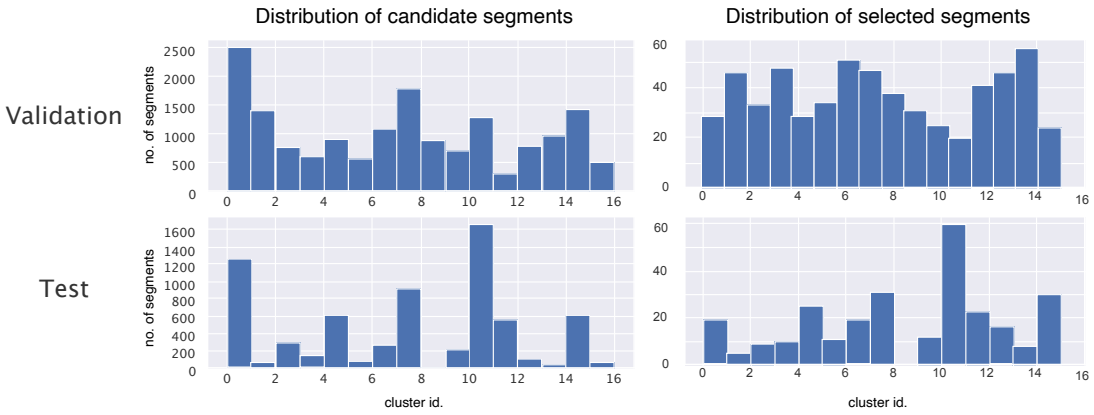

Figure 5: Distribution of the candidate segments (left column) and the sampled final segments (right column) from the validation and test sets (top and bottom rows, respectively).

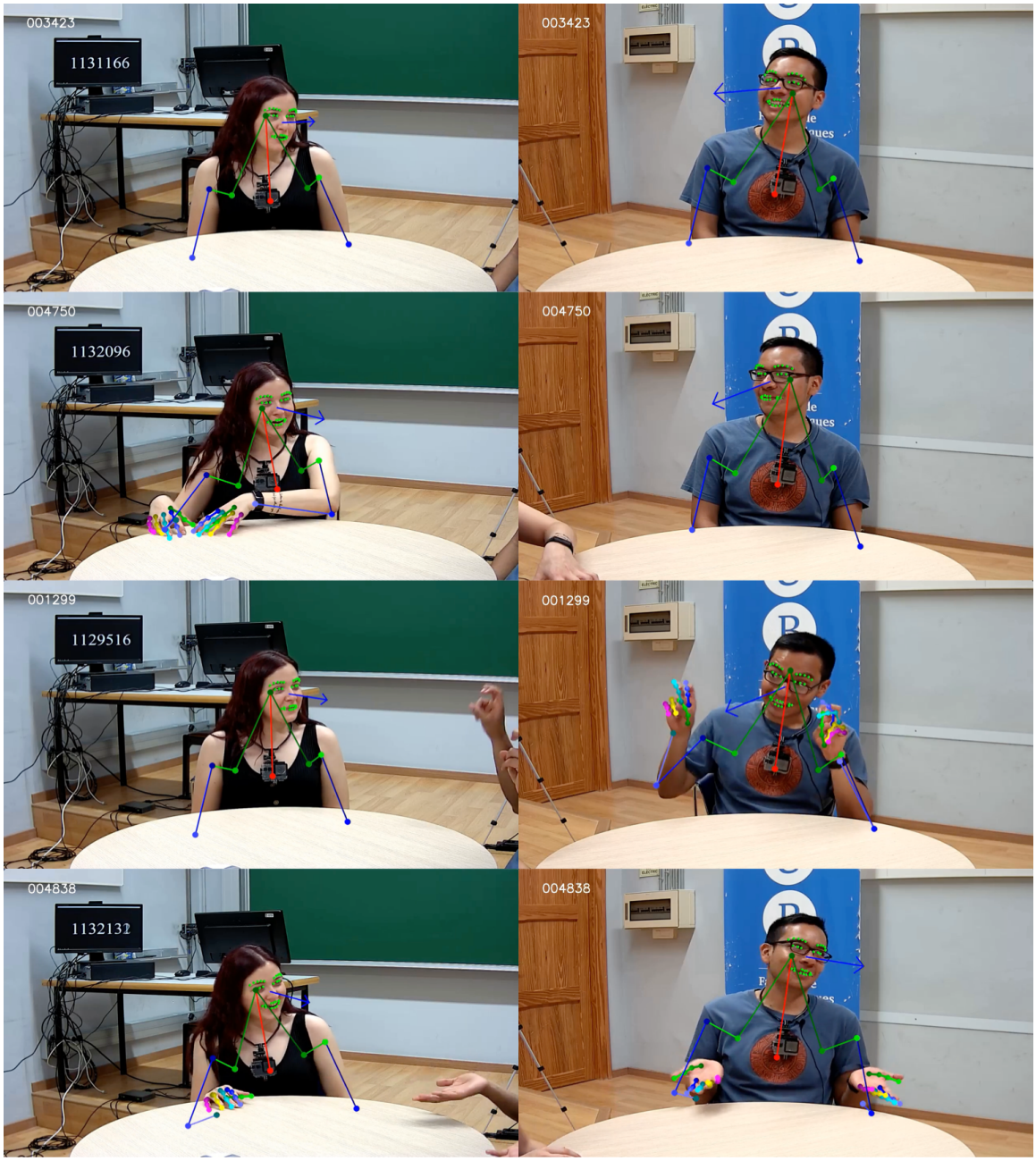

Figure 6: Examples of behaviors featured in segments associated to clusters 1, 8, 10, and 11, from top to bottom. The selected clusters capture behavioral patterns where hands are either hidden for both participants (top row), visible for at least one of them (mid rows), or visible for both (bottom row).