UDIVA V0.5 Dataset

The UDIVA v0.5 dataset is a preliminary version of the UDIVA dataset, including a subset of the participants, sessions, synchronized views, and annotations of the complete UDIVA dataset. The UDIVA v0.5 dataset is composed of 145 dyadic interaction sessions divided in 4 different tasks each: *Talk*, *Lego*, *Ghost*, and *Animals*. Such sessions are performed by 134 participants (ranging from 17 to 75 years old, 55.2% male), who can participate in up to 5 sessions with different participants.

The UDIVA v0.5 dataset consists of the subset of recordings and metadata used for the evaluation of the context-aware personality inference method presented here. In addition to that data, we are adding the transcripts of the session conversations, and a set of automatically extracted annotations, namely:

- Face landmarks: 68 face fiducials were extracted using the 3DDFA_v2 algorithm, and the detection confidence provided by the face detector (Faceboxes). Additionally, an average smoothing with the immediate previous and next frames was applied.
- Body landmarks: full body joints and a detection confidence were extracted using the MeTRAbs method.
- Hand landmarks: hand landmarks and detection confidences were extracted by FrankMocap. Additionally, several post-processing steps were applied to them to improve their quality:
    1. Hand detections needed for the landmark extractions were tracked with SiamRPN++ when temporal or big spatial gaps were found.
    2. Body pose and hand landmarks were leveraged to ensure that only the hands from the person of interest were detected.
    3. For the errors identified, a second landmark extraction was run.
- 3D eye gaze vectors: gaze vectors were extracted using ETH-XGaze, using the previously extracted face landmarks and a dummy camera matrix.

Validation and test sets underwent visual inspection in order to assess their accuracy. For each frame, raters manually checked face, body and hands landmarks (the gaze vector was not assessed):

- Face landmarks were discarded (*valid* flag set to *False*) when either the face orientation or the landmarks position was slightly wrong.
- Body landmarks were discarded (*valid* flag set to *False*) when either the overall pose was considered mistaken or one side was strongly displaced from the real joint positions.
- Hand landmarks:
    - Landmarks were discarded (*valid* flag set to *False*) when fingers were strongly displaced. Mild fingers displacements were tolerated when the overall orientation and hand placement was correct.
    - Landmarks of hands hidden under the table (false positives) were discarded.

- Mismatched hands with switched left/right labels were swapped.
- For frames within periods of time (t_0 to t) in which the participant did not move their hands but predicted landmarks were wrong (e.g. due to hands interaction), landmarks were interpolated using those from frames t_0*-1* and *t+1*, if the rater considered that such interpolation would yield landmarks fulfilling our accuracy standards.

Data structure and annotations

The UDIVA v0.5 Dataset is divided into 3 subject-independent splits, as defined here: train (116 sessions and 99 participants), validation (18 sessions and 20 participants), and test (11 sessions and 15 participants). Sessions are identified with a 6-digit string identifier, while participants are identified with an integer identifier. Tasks are identified with their initial letter or with their full name (T - talk, G - ghost, A - animals, L - lego). The recording of each task starts when the task administrator finishes explaining the task to the participants, and stops when the administrator starts interacting with the participants again to deliver the following task. The real task (for instance, build a Lego building) may finish way before the end of the recording, after that participants are free to continue playing or just wait until the task administrator enters the recording room and stops the recording.

The dataset is organized with one folder per split. Within each split folder, there is one folder per task, such that users can decide which task they want to use for their research and download just that task, without having to download the whole dataset.

Note that for the behavior forecasting track of the DYAD@ICCV2021 challenge, challenge participants are requested to predict the landmarks of given segments of the 'talk' task for validation and test sets. Therefore, specific recording and transcription segments of the 'talk' task in validation and test sets will be masked out; that is, no image, audio, nor transcripts will be provided for the segments to be predicted. As both behavior forecasting and personality recognition tracks use the same data, researchers participating in the personality recognition track will also be provided with the same masked data.

We describe the structure of each split below:

Audio-visual recordings (recordings folder)
Audio-visual recordings, in .mp4 format. There is one recording per task and session. Within each folder, there are 2 .mp4 video files per task with name format FCX_Y.mp4, X denoting the participant view (1 or 2) and Y denoting the task.

Transcripts (transcriptions folder)
There is one folder per session, containing one *.srt* file per task included in that session, with name format SESSIONID_Y.srt where Y denotes the task. In the transcription files, each participant is identified as *PART.1* (visible from FC1 view) or *PART.2* (visible from FC2 view). If a

given task of a given session does not have transcriptions, there will be no file for that session and task.

Annotations (annotations folder)

There is one folder per session and one subfolder per subject. Each subfolder contains a compressed .hdf5 file ("annotations_raw.zip"), containing automatically extracted landmarks for hands, body, and face, and 3D eye gaze vectors. Landmarks of segments belonging to validation and test sets to be predicted in the DYAD challenges are not provided to the challenge participants.

The hdf5 file contains a "group" structure per frame (e.g., "00010" for frame #10). Within each group we have the following information (italic for *attributes structure* and bold for **dataset structure**):

- valid: boolean. Whether the frame is valid or not.
- face: "group" structure:
    - *valid*: Boolean. Only available in the validation and test sets. It indicates if the landmarks provided are correct according to the precision standards established.
    - *gaze*: array of 3 floats representing the 3D gaze vector. If missing, it was not correctly extracted.
    - *confidence*: provided by the face detector.
    - **landmarks**: array of size (68, 3) of int16 values (68 landmarks, 3 coordinates)
- body: "group" structure:
    - *valid*: Boolean. Only available in the validation and test sets. It indicates if the landmarks provided are correct according to the precision standards established.
    - *confidence*: provided by the body detector.
    - **landmarks**: array of size (24, 3) of int16 values (24 landmarks, 3 coordinates).
- hands: "group" structure:
    - left (and same structure for 'right'): "group" structure:
        - *valid*: Boolean. Only available in the validation and test sets. It indicates if the landmarks provided are correct according to the precision standards established.
        - *visible*: Boolean. For the training set, it indicates whether the hand was detected. In the validation and test sets, whether the hand is inside the field of view or not. Please, mind the scenario where the hand

may have been detected but landmarks were not extracted successfully.
    - *confidence*: provided by the hands detector. If '-1', the hand bounding box has been tracked as part of the post-processing.
    - **landmarks**: array of size (21, 3) of int16 values (21 landmarks, 3 coordinates).

Note: for the cases where the extraction failed, the landmarks dataset is not available.

Metadata (metadata folder)

Metadata_<split>.zip, contains csv files for participant and sessions metadata:

- Parts_<split>.csv: participants metadata file, with one participant per row, including the following information:
    - participant ID,
    - self-reported gender (F - female, M - male),
    - self-reported age (integer),
    - country of origin (string),
    - max. education level (string),
    - self-reported personality questionnaire (BFI-2) results, as *OPENMINDEDNESS_Z, CONSCIENTIOUSNESS_Z, EXTRAVERSION_Z, AGREEABLENESS_Z, NEGATIVEEMOTIONALITY_Z* (float, 0-centered) (personality values for validation and test sets are not provided to the DYAD challenge participants).
    - total number of sessions done (integer),
    - session IDs in which the participant has participated in order of occurrence (*SESSION1* to *SESSION5*, including only sessions included in UDIVA V0.5, string).

- Session_<split>.csv: sessions metadata file, with one session per row, including the following information:
    - session ID (string),
    - participant ID corresponding to *PART.1* (string),
    - participant ID corresponding to *PART.2* (string),
    - recording timestamp (datetime),
    - difficulty level of Lego task (integer, from 0 to 3),
    - difficulty level of Animal task (integer, from 0 to 2),
    - language (string),
    - notes (string),
    - self-reported relationship among participants (0 - they didn't know each other, 1 - they knew each other),
    - order of administration of the tasks within a session (*X_ORDER,* where X is the task name - *talk, lego, ghost, animals,* integer from 1 to 4).

- animal assigned to each participant in the Animals task (*ANIMAL_PART.1/2,* string).
- self-reported answers of the mood questionnaire (PEQPN) per participant (1 and 2) before and after the session (*PEQPN_GOOD_X_Y, PEQPN_BAD_X_Y, PEQPN_HAPPY_X_Y, PEQPN_SAD_X_Y, PEQPN_FRIENDLY_X_Y, PEQPN_TENSE_X_Y, PEQPN_RELAX_X_Y*, where X is *BEFORE* or *AFTER*, and Y is *PART.1* or *PART.2,* all integers from 1 to 5),
- self-reported fatigue levels per participant (1 and 2) before and after the session (FATIGUE_X_Y, where X is *BEFORE* or *AFTER*, and Y is *PART.1* or *PART.2,* all integers from 0 to 10 - the questionnaire contained values from 1 to 10, participants that did not answer this question have a value of 0 assigned),

task_limits.json file: A .json file with the start-end times (in seconds) of the actual task at hand (e.g., the time when participants start building the Lego figurine and when they finish doing so), per session and task, with format:

```
{"session_ID1": {
   "talk": [[start, end]],
   "ghost": [[start, end]],
   "lego": [[start1, end1],[start2,end2]],
   "animals": [[start, end]],
   },
 "session_ID2": {
   ...
   },
 ...
  }
```

Notice the case of “lego” in the example above. If a recording was interrupted for a given reason (explained in *defective_segments.json* below and in the README file provided with the dataset), multiple time limits will be given, 1 list per each valid segment. In case any of the tasks are not available for a given session, those tasks will contain an empty list.

defective_segments.json file: .json file that describes particularities of the sessions/tasks, if any. This may include: start-end timestamps (in seconds) of segments that do not contain audio (*no_audio*) or video (*no_video*); start-end timestamps (in seconds) for invalid video segments (*invalid*, i.e., black segments in the video with neither image nor audio, which are removed from the data due to privacy reasons). The file will only contain as keys the session IDs, task IDs, and view IDs of those sessions, tasks, and views that contain a particularity. Example of the format:

```
{"session_ID1":
 {
   "lego": { "invalid": [[50.2,58.1], [144.9, 150.1]] },
  }
 "session_ID40":
```

```
{
  "talk" : {
        "no_audio": {
                    "FC1_T":[[0.0,180.0]],
                    },
        },
}
...
}
```

Participants consent

sharing_consents.csv file: .csv file, with one participant per row, specifying whether the participant gives consent to researchers to use their image/audio/transcript for illustration purposes in publications only, as per the Dataset License. It includes the following information:

- participant ID (string),
- consent to use images (*X* denotes confirmed consent),
- consent to use audio (*X* denotes confirmed consent),
- consent to use transcripts (*X* denotes confirmed consent).